ets research institute

SEPTEMBER 2026
RR–26-10
TOEFL-RR-110

TOEFL® RESEARCH SERIES

Linguistic Features in Young English-as-an-Additional-Language Learners' Writing: A Classroom-Based Longitudinal Study

AUTHORS

Ching-Ni Hsieh and Huiying Cai

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

Daniel F. McCaffrey
Lord Chair in Measurement and Statistics

ASSOCIATE EDITORS

Usama Ali
Principal Measurement Scientist

Beata Beigman Klebanov
Principal Research Scientist, Edusoft/ETS

Katherine Castellano
Managing Principal Research Scientist

Michael Fauss
Research Scientist

Yang Jiang
Research Scientist

Jamie N. Mikeska
Managing Principal Research Scientist

Teresa Ober
Research Scientist

Jonathan Schmidgall
Senior Research Scientist

Jesse Sparks
Director Research

Zuowei Wang
Senior Research Scientist

Mikyung Wolf
Principal Research Scientist

Diego Zapata Rivera
Distinguished Presidential Appointee

Klaus Zechner
Senior Research Scientist

Jiyun Zu
Senior Measurement Scientist

PRODUCTION EDITORS

Ayleen Gontz
Manager Editorial Services

Emily Ramsower
Manager Editorial Services

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

Linguistic Features in Young English-as-an-Additional-Language Learners' Writing: A Classroom-Based Longitudinal Study

Ching-Ni Hsieh[1] & Huiying Cai[2]

[1]Center for English Language Learning and Assessment, ETS Research Institute, ETS, Princeton, New Jersey, United States

[2]Department of Linguistics, University of Illinois Urbana-Champaign

Abstract

This longitudinal study examines patterns of linguistic changes in young English-as-an-additional-language (EAL) learners' written responses across tasks, score levels, and test administrations over a 1-year period. Writing samples were collected from 114 learners ages 8 through 14 using a descriptive task and an email task. The written responses were scored holistically by trained raters, and five linguistic features were extracted using automated text analysis tools. Results showed that higher scoring responses were generally longer, exhibited greater lexical diversity, included words acquired later in life, contained less imageable vocabulary, and demonstrated greater syntactic variety. Task type also influenced the response characteristics. The email responses tended to show greater linguistic complexity than the descriptive responses, particularly at the lower score levels, although the differences between tasks diminished at higher score levels. Analyses across administrations revealed systematic differences in response characteristics over time, with larger changes observed between the first and second administrations than between the second and third administrations. Among the linguistic features examined, syntactic variety showed the strongest association with both score differences and temporal variation, whereas lexical sophistication measures displayed more nuanced and nonlinear patterns. These findings highlight score-, task-, and time-sensitive variations in young EAL learners' written responses and the importance of examining different linguistic dimensions when characterizing young learners' writing quality in classroom-based longitudinal data.

Keywords: young learners, English as an additional language, longitudinal study, writing tasks, email, description, lexical diversity, syntactic complexity

Corresponding author: Ching-Ni Hsieh, E-mail: chsieh@ets.org

Introduction

Differences in writing quality are frequently examined through changes in linguistic features, such as lexical diversity and syntactic complexity, exhibited in language learners' written texts. These linguistic features are commonly used as criteria for evaluating language learners' written texts and describing performance levels (Bulté & Housen, 2014) and can serve as benchmarks for tracking language development (Kim, 2021; Kyle et al., 2021).

Extensive research has explored the relationships between linguistic complexity and writing quality in second language (L2) learners' texts (e.g., Kyle & Crossley, 2016; Kyle et al., 2021; Taguchi et al., 2013; Yang et al., 2015). This body of work has yielded valuable insights into L2 writing quality and informed both instructional and assessment practices. However, most of these studies rely on cross-sectional data from college-level or adult L2 learners (e.g., Biber et al., 2004; Lu, 2011). In contrast, relatively few studies have focused on young English-as-an-additional-language (EAL) learners (e.g., De Wilde, 2023; Hsieh, 2024) and the linguistic characteristics in their writing (Sung et al., 2025; Wolf et al., 2018). *Young EAL learners* here refer to school students ages 6 through 16 , an age range that includes primary and lower secondary schooling in many EAL learning contexts (e.g., Loder Buechel, 2015; Patekar, 2021; Vogt et al., 2020).

The majority of studies that explore young EAL learners' writing quality or linguistic changes utilize cross-sectional data (e.g., Crossley et al., 2011; De Wilde, 2023; Levitzky-Aviad & Laufer, 2013; Michel et al., 2019; Wolf et al., 2018). These studies generally find that older children outperform the younger ones in their overall writing performances (e.g., Michel et al., 2019) and higher scoring texts tend to include more sophisticated words and more complex sentence structures (e.g., Crossley et al., 2011; Wolf et al., 2018). Despite this general trend, linguistic changes and children's writing development are not always linear and occasional decrease in linguistic complexity is observed in children's written texts (e.g., Levitzky-Aviad & Laufer, 2013).

Although previous cross-sectional studies offer meaningful information about the general trends of linguistic changes in young EAL learners' written texts, longitudinal research is

essential for capturing linguistic changes over time (Crossley et al., 2016; Kyle et al., 2021). Thus far, however, longitudinal investigations of young EAL learners' writing quality and linguistic variations remain limited (e.g., Gené-Gil et al., 2015; Levitzky-Avid & Laufer, 2013; Wang, 2022), leaving a critical gap in our understanding of how writing skills evolve during the early stages of second language acquisition (SLA). Building on prior research and gaps in the literature, our study examines linguistic changes in young EAL learners' written texts across task types and score levels and over time.

Lexical Complexity

Lexical complexity, typically measured by lexical density (e.g., proportion of content words in a text), diversity (use of different types of words), and sophistication (e.g., word frequency), is a key indicator of writing quality (Bulté & Housen, 2014). Higher quality texts tend to exhibit less frequent words, words acquired at a later age, and more diverse words (Crossley, 2020).

Longitudinal studies involving adult L2 learners have shown that lexical diversity and sophistication generally increase with better writing quality (e.g., Kim, 2021; Mazgutova & Kormos, 2015; Zheng, 2016), although studies conducted over short time spans (e.g., one semester) often report minimal quality change (e.g., Bulté & Housen, 2014; Yoon & Polio, 2017). Similar trends on the longitudinal development of lexical properties are observed in the writing of young EAL learners (e.g., Gené-Gil et al., 2015; Maamuujav, 2021). For example, Gené-Gil et al. (2015) examined writing development among 50 students in Grades 8 to 10 at four high schools in Spain. The students completed a timed email task (25 minutes) at the beginning of the second, end of the second, end of the third, and end of the fourth academic years. Their essays were analyzed for lexical and syntactic complexity, accuracy, and fluency. Using D-value, a measure of vocabulary diversity developed by Malvern and Richards (1997), the researchers noted that students produced texts with increasing lexical variation over time with significant improvements between different time intervals.

Syntactic Complexity

Syntactic complexity refers to the variation and sophistication of grammatical structures in writing (Ortega, 2015). In SLA and assessment research, it is commonly measured using both global features, such as mean sentence length, and more fine-grained features of clausal and

phrasal complexity (Bulté & Housen, 2017; Kyle & Crossley, 2018). These features have been examined extensively in the writing of adult L2 learners (e.g., Kim, 2021; Mazgutova & Kormos, 2015).

Longitudinal research on young EAL writers is still emerging. Wang (2022) conducted a case study and analyzed diary entries written by an 11-year-old Chinese EAL learner over 1 year. Using the L2 Syntactic Complexity Analyzer (Lu, 2010), the researcher computed various sentential, clausal, and phrasal complexity measures of the compiled written texts. The study findings revealed a dynamic nonlinear pattern in the writer's syntactic complexity; some measures (e.g., mean sentence length, mean T-unit length) showed steady increases while others (e.g., prepositional phrases, relative clauses) fluctuated. In another longitudinal study, Gené-Gil et al. (2015) found that secondary students in a content and language integrated learning (CLIL) program produced increasingly complex texts over 3 years, particularly in their use of coordination and subordination.

Young EAL Learners' Writing Quality and Linguistic Features

Few studies have examined the relationship between linguistic features and overall writing quality in young EAL learners' texts (e.g., Jiang et al., 2019; Maamuujav, 2021; Martínez, 2018). Using cross-sectional data, Maamuujav (2021) explored a range of lexical features and their relationship with writing quality in academic essays written by seventh- to 12th-grade English learners in U.S. schools. Findings revealed that lower grade students' essays tended to have lower lexical density (lower percentage of content words and higher percentage of function words) and diversity and that low-scoring essays contained many basic and high-frequency words used in everyday communication. Additionally, lexical sophistication, based on age of acquisition measure and the percentage of academic words, significantly predicted holistic writing scores. In another study, De Wilde (2023) analyzed a learner corpus that contains narrative essays written by 3,168 EAL students, 11 and 12 years old, in Belgium. The analysis indicated that lexical measures of word frequency, age of acquisition, and imageability significantly predicted students' holistic essay scores. This result is in line with findings of other studies (e.g., Maamuujav et al., 2021) that show more proficient young EAL learners tend to use less imageable and more sophisticated words in their written texts.

Research on the links between syntactic complexity and young EAL students' written texts suggests that higher quality writing contains more complex syntactic features. For example,

Jiang et al. (2019) analyzed 410 narrative essays written by young Chinese EAL students. Findings indicated that measures such as mean T-unit length, mean sentence length, and dependent clauses per clause significantly differentiated essays across proficiency levels. Similarly, Martínez (2018) analyzed writing samples from 188 Spanish EAL learners who completed a timed writing task (30 minutes) in the classroom. The written texts were scored holistically and coded on a range of syntactic complexity measures. The analysis revealed significant and strong correlations between holistic essay scores and syntactic complexity features such as mean sentence length and coordinate clause ratios.

Two empirical studies used cross-sectional data collected from a standardized writing assessment to explore the relationship between young EAL learners' writing performance and linguistic characteristics (Sung et al., 2025; Wolf et al., 2018). Sung et al. (2025) examined lexical richness indices in relation to scores on two TOEFL Junior® Opinion and Listen–Write tasks. Results showed that students' lexical characteristics differed by task score and task type, particularly in lexical diversity and sophistication. Also using TOEFL Junior Writing test data, Wolf et al. (2018) analyzed responses to the TOEFL Junior Opinion writing task and found clear relationships between various lexical and syntactic features and holistic task scores assigned by human raters.

Research Questions

Young EAL students differ cognitively from adult language learners, and their linguistic changes exhibited in their written texts may follow distinct trajectories (Butler, 2025; Pinter, 2006). Despite this, as noted earlier, young EAL learners' linguistic characteristics and writing quality remain underexplored. A better understanding of the linguistic characteristics exhibited in young learners' writing and their changes over time can offer insight into writing skills of young EAL learners and help explain linguistic variations from a language acquisition angle. The results can also inform classroom teachers about young learners' linguistic needs and language skills that are necessary to meet the demands of English writing in formal classroom instruction.

The current study examined linguistic differences in young EAL learners' written texts across score levels, task types, and test administrations. The research questions guiding this study included the following:

RQ1. What is the relationship between linguistic features and holistic task scores of young EAL learners' descriptive and email essays?

RQ2. How do the linguistic features in young EAL learners' descriptive and email essays change over the course of 1 year?

Methods

Context

This study was conducted as part of a broader research project led by a team of researchers at ETS in collaboration with language educators from The Netherlands, Denmark, and Finland. The larger project, titled Innovating and Building Understanding to Improve Language Development with Information Technology (iBUILD IT), was funded by an ERASMUS+ grant from the European Union. The iBUILD IT project aimed to (a) develop innovative and sustainable solutions for delivering high-quality English language education and (b) bridge the gap in primary and secondary English language education by designing practical guidelines to support students during this transition (Erasmus, 2022). To achieve these goals, the project team organized a series of face-to-face teacher training workshops and online webinars to enhance teachers' pedagogical approaches and language assessment literacy. Concurrently, the ETS research team undertook independent studies to examine EAL students' writing and speaking skills and the challenges faced by young learners transitioning from primary to secondary education.

Within the context of the iBUILD IT project, based on information provided by the classroom teachers, writing is typically taught from the second or third grade level and a communicative approach to writing is generally promoted across the three EAL contexts. English language curriculum and textbook series such as the *Look* published by National Geographic (https://www.eltngl.com/series/PRO0000009082) and *Stepping Stones* by Noordhoff (https://www.noordhoff.nl/voortgezet-onderwijs/engels/stepping-stones), along with teacher-designed materials, are used for instruction. Classroom writing activities vary from various closed, controlled exercises (e.g., fill in the blank, sentence completion) to more open, modeled writing tasks (e.g., paragraph completion, personal correspondence, short essay) depending on students' grade level and proficiency and the local curriculum and guidelines.

Students and Samples

The iBUILD IT project involved students from three primary and three lower secondary schools located in Denmark, Finland, and The Netherlands. The project spanned two academic years from Fall 2021 to Spring 2023, engaging approximately 600 students. The students' grade levels ranged from third to sixth grades in primary schools and seventh and eighth grades in lower secondary schools. The age of the participants varied from 8 through 14 years (M = 12.19, SD = 1.42). Most students spoke Danish, Finnish, or Dutch as their first language. No other demographic information was collected or reported here to protect students' privacy.

In this study, we focused on a subgroup of students who completed two writing tasks at three waves of data collection (Times 0–2) between Fall 2021 and Fall 2022 (see the Data Collection section). This subgroup consisted of 114 students, including 27 primary school students (Grades 4 to 6) and 87 lower secondary school students (Grades 7 and 8).

Writing Tasks

Students' linguistic characteristics and writing quality were assessed through two tasks: one descriptive task and one email task (see Figures 1 and 2). The descriptive task was developed by a team of researchers and assessment developers at ETS. The task required students to write a short paragraph about a familiar topic in 15 minutes on the computer. The topics varied across the three waves of data collection, including describing (a) an experience playing with friends at Time 0, (b) a place you want to visit at Time 1, and (c) a school activity you enjoy doing at Time 2. The email task was adapted from one of the TOEFL Junior Writing assessment tasks[1] (ETS, 2025a). To accommodate the younger age group of some participants (younger than the typical TOEFL Junior Writing test takers ages 11 and above), in consultation with the classroom teachers, we extended the response time from 7 to 15 minutes. In this task, students read an email about a school club and responded to three embedded questions. The school clubs varied and included (a) chess club (Time 0), (b) basketball club (Time 1), and (c) music club (Time 2).

Because the classroom teachers followed a standardized task administration manual to guide the students and implemented the tasks within their regular English classes in this longitudinal, classroom-based research with young learners across three countries, a counterbalanced design that rotated tasks and topics across students and waves was not feasible. To minimize prompt-related variability and enhance comparability across administrations, we took several steps during test development and implementation. First, the selection of the

prompts was guided by our close collaboration with the classroom teachers through multiple project meetings and face-to-face workshops. We chose topics that were age appropriate, familiar, and broadly accessible to students across contexts. These topics reflected everyday experiences and school-related themes about which young EAL learners were likely to have sufficient background knowledge to complete a writing task. In addition, readability analyses provided quantitative evidence of prompt comparability. The Flesch-Kincaid readability scores (FK scores; Flesch, 1948) for the three descriptive prompts were 88, 87, and 80, all falling within the same score category indicating that the texts were easy to read. The FK scores for the three email prompts were 97, 92, and 95, all falling within the same score category indicating very easy readability. These indices suggested that the prompts across waves were comparable in linguistic difficulty and unlikely to systematically advantage or disadvantage students at particular time points. Finally, across all administrations, the two task types followed identical task structures, time limits, response requirements, delivery modes, and scoring criteria.

Figure 1. Sample Descriptive Task

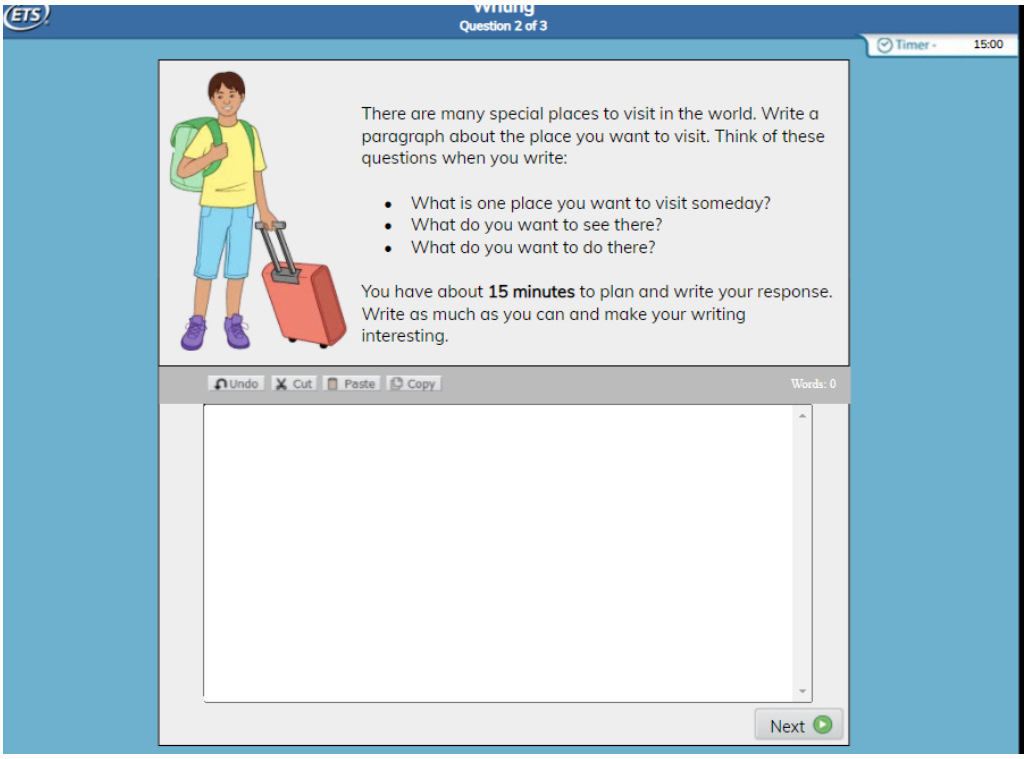
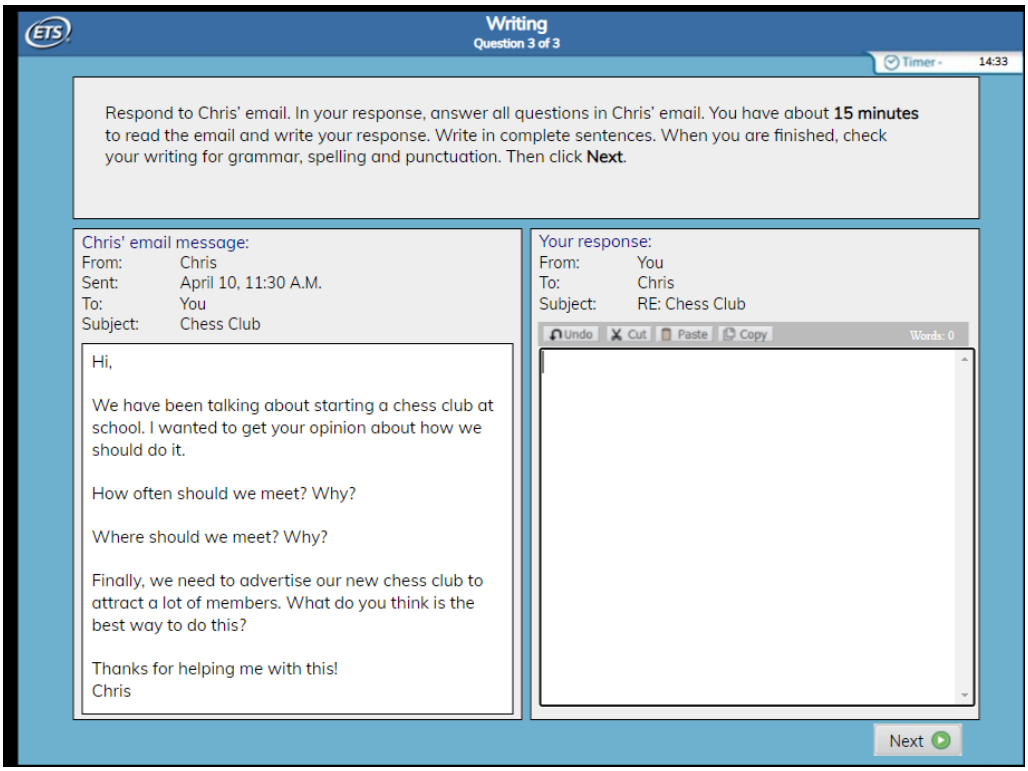


Figure 2. Sample Email Task



Data Collection

Data was collected in three waves: October/November 2021 (Time 0), April/May 2022 (Time 1), and October/November 2022 (Time 2), with approximately 6 months between each wave. The classroom teachers administered the writing tasks to their students during regular English classes. The teachers were provided with a detailed, step-by-step administration manual to guide the students, ensuring that they understood the topics and task requirements and that technical issues, if any, were resolved. Teachers reported that the students were familiar with reading and typing in English on the computer and that the test administrations went smoothly.

Data Analysis

Scoring of Written Responses

A total of 342 descriptive essays and 342 email essays produced by the 114 students were analyzed to examine their writing quality and linguistic changes over the 1-year data collection period. The essays were scored by two project team members with graduate degrees and research experience in language education and assessment. They evaluated the overall quality of the

responses using a 4-point holistic scale that included performance criteria on language use and content development. These criteria were consistent with components later adopted in scoring the TOEFL Primary® and TOEFL Junior Writing tests (ETS, 2022b, 2023), which were launched after the current data were collected. To achieve the maximum score of 4 points, a written response needs to demonstrate appropriate vocabulary and sentence structures, a logical progression of ideas with minimal unrelated content, and good use of transitions.

The two raters completed several rounds of training and norming sessions. They double-rated a random sample of 118 essays to establish interrater reliability using Cohen's weighted kappa (k = .81), indicating substantial agreement. All discrepant cases were discussed until consensus scores were reached. The remaining texts were scored by one rater. To ensure consistency of scoring, the raters met periodically to recalibrate when necessary and discuss difficult cases throughout the rating process.

Analyses of Linguistic Features

The same two raters assessed the students' language use in their essays, using a 4-point scale that was developed as part of the iBUILD IT project. Language use was operationalized as the overall accuracy and variety of vocabulary and sentence structures demonstrated in the essays. To ensure consistency, the raters went through multiple training and norming sessions. They double-rated a random sample 205 essays to establish rater reliability with k = .86, indicating a high degree of agreement. All discrepant cases were resolved until consensus scores were reached. The remaining essays were evaluated by one rater.

To analyze linguistic changes, we utilized two natural language processing (NLP) text analysis tools, e-rater® and TextEvaluator®, to extract automated linguistic features from the students' essays. The e-rater tool, originally developed for automated essay scoring, analyzed essays for linguistic accuracy (e.g., lexical and grammatical errors) and complexity (e.g., lexical density, syntactic variety; Attali & Burstein, 2006). Building on the e-rater system, TextEvaluator included additional automated lexical and syntactic complexity measures such as mean sentence length, dependent clause count, age of acquisition (Bird et al., 2001; Kuperman et al., 2012), or word imageability (Bird et al., 2001; Coltheart, 1981), which were not available in e-rater (Sheehan et al., 2010).

More than 100 textual features were generated from the e-rater and TextEvaluator tools. To narrow the scope for the current study, drawing on prior research in L2 writing, we selected a

preliminary list of 20 NLP-based linguistic features for our initial investigation. These features included measures of essay length, lexical diversity, lexical sophistication, and syntactic complexity and variety. To identify features most closely associated with young EAL writers' linguistic performance, we subsequently conducted a correlation analysis between the preliminary list of features and human-rated language use scores. Given that language use score was a categorical variable, we examined Spearman's rho correlations between each NLP-based feature and language use score and identified a small set of features with meaningful correlations with the scores ($|r| > .3$; Hinkle et al., 2003). Among this set of features, we further examined Pearson intercorrelations to filter out redundant features. High intercorrelations ($|r| > .8$) indicated that feature pairs measured a similar construct (Young, 2017). In such cases, we excluded the feature with the lower correlation with language use scores.

After finalizing the set of linguistic features, we converted them to z-scores and analyzed them using linear mixed-effects models (LMMs) to examine how these features varied across essay scores and changed over time yet still accounted for repeated measures from the same students. In addition, we explored whether young EAL learners' writing performances differed by task type and education level—specifically, primary versus lower secondary education. We were specifically interested in identifying similarities and differences in writing characteristics and developmental patterns between primary and secondary students to inform classroom instruction.

In this study, separate LMMs were fitted for each linguistic feature, with "student" treated as a random effect. We treated the retained linguistic features as dependent variables. For each model, we examined random intercepts using intraclass correlation coefficient (ICC). All ICCs indicated meaningful variations in every dependent variable associated with the responses across individual students (ICC > .05; Snijders & Bosker, 2011). As a result, individual students were included as random intercepts in all models.

We subsequently selected fixed effects for each dependent variable (i.e., linguistic feature). To address RQ1, we included the following independent variables in our analysis: essay scores, writing tasks, and education levels. For each model, we added these variables one at a time into the model as a fixed effect and used likelihood ratio tests (LRT) to examine whether each addition significantly improved the model. Variables that did not yield statistically significant improvements were excluded.

When adding a variable, we also examined whether adding a random slope was necessary. A random slope was only included if the addition led to significant model improvement without causing overfitting. Once the fixed effects were determined, we tested for fixed interaction effects among the selected variables one at a time using LRT. Interactions that resulted in significant model improvements were retained.

To answer RQ2, we examined the three independent variables: time, writing tasks, and education levels. The fixed effects were selected following the same process used for RQ1. For each final model, we performed post-hoc analyses using Tukey's Honestly Significant Difference (HSD) tests to perform pairwise comparisons.

Results

RQ1. What Is the Relationship Between Linguistic Features and Holistic Task Scores of Young EAL Learners' Descriptive and Email Essays?

The correlation analysis revealed that five out of the initial list of NLP-based linguistic features had meaningful relationships with human-rated language use scores (LG scores; $|r| > .3$) while reflecting distinct constructs (i.e., $|r| = .12$ to .68). These features were essay length, lexical diversity, age of acquisition, imageability, and syntactic variety score. They were included in the final set of linguistic features for subsequent analysis (see Table 1 for descriptions).

Table 1. Descriptions of the Final Set of Linguistic Features

Feature	Description	Natural language processing tool
Essay length	Total number of words	e-rater®
Lexical diversity	Type-token ratio adjusted for text length using the formula: Type count divided by the square root of twice the word count	TextEvaluator®
Age of acquisition	Average age of acquisition score of all words	TextEvaluator®
Imageability	Average imageability score of all words	TextEvaluator®
Syntactic variety	A score assessing whether writers demonstrate control over a wide range of syntactic features of the language	e-rater®

It should be noted that both age of acquisition and imageability are psycholinguistic measures of words in a text that are used as indicators of lexical sophistication and have shown to predict writing quality in extant studies (e.g., Kim, 2021; Kyle & Crossley, 2016; Maamuujav,

2021). An age of acquisition score indicates how early or late a word is learned, with lower scores representing earlier acquisition and higher scores indicating later acquisition. Words with low age of acquisition scores are processed faster and more accurately by adults, and words with high scores are processed more slowly. A word's imageability score indicates how easily it evokes a mental image, sound, or other sensory experience, with higher imageability scores for more concrete words such as *houses* and lower scores for more abstract words such as *justice*.

As shown in Table 2, language use scores were significantly positively correlated with essay length, lexical diversity, age of acquisition, and syntactic variety score but negatively correlated with imageability. These correlations suggest that higher scoring responses tended to be longer, include more diverse and less imageable words, incorporate words acquired later in life, and contain more varied sentence structures.

Table 2. Correlations Between Linguistic Features and Language Use Score (LG Score)

Linguistic feature	0[a]	1	2	3	4	5
0. LG Score	—					
1. Essay length	.52*	—				
2. Lexical diversity	.59*	.68*	—			
3. Age of acquisition	.39*	.23*	.37*	—		
4. Imageability	-.32*	-.28*	-.42*	-.12*	—	
5. Syntactic variety	.46*	.64*	.66*	.30*	-.34*	—

Note. Numbers in the row heads correspond to the column heads.
[a] The correlations with LG scores were Spearman's rho correlation coefficients.
* $p < .05$

Table 3 presents descriptive statistics of the five linguistic features across essay scores. Essays with higher scores tended to have greater length, higher lexical diversity, words acquired later in life, greater syntactic variety, and lower levels of imageability. This trend aligns with the relationships between the linguistic features and language use scores described previously.

Table 3. Descriptive Statistics of Linguistic Features for Each Essay Score

Essay score	*n*	Mean (*SD*)				
		Essay length	Lexical diversity	Age of acquisition	Imageability	Syntactic variety
1	74	37.27 (27.37)	2.98 (0.71)	2.19 (0.78)	3.57 (0.33)	1.31 (0.64)
2	255	66.05 (37.40)	3.57 (0.66)	2.42 (0.79)	3.50 (0.21)	1.67 (0.57)
3	220	114.42 (54.89)	4.26 (0.56)	2.73 (0.41)	3.44 (0.16)	2.19 (0.53)
4	135	129.82 (60.38)	4.51 (0.65)	2.78 (0.32)	3.40 (0.13)	2.35 (0.45)

Note. Number of individual students = 114; total *N* = 684.

For each linguistic feature, LMMs were selected following the process described in the Methods section to examine variations across essay scores, tasks, and education levels. All final models met the LMM assumptions, except for the model predicting essay length. Residual plots revealed that the essay length model violated the homoscedasticity assumption. To address this issue, we square-root-transformed essay length and standardized it, which then satisfied the model assumption.

Table 4 reports the results of the final LMMs for each linguistic feature. Education level was excluded from the final models because its inclusion did not significantly improve model performance. The reference levels used in the LMMs were an essay score of 1 and the descriptive task. The results indicated that both essay scores and task types had significant main effects on the five linguistic features. However, neither effect had a random slope, suggesting no meaningful individual-level difference in how essay scores and task types affected the linguistic features.

As shown in Table 4, email responses, compared to descriptive responses, featured more diverse vocabulary, a greater proportion of words acquired later in life, less imageable words, and more varied sentence structures while maintaining similar essay lengths. Despite these differences, higher scoring responses were generally longer and demonstrated more diverse vocabulary, more words acquired later in life, less imageable words, and more syntactic variety compared to lower scoring essays. These results suggested that email responses tended to exhibit greater linguistic complexity than descriptive responses and that higher scoring responses, in general, were characterized by more complex lexical and syntactic features.

According to the effect sizes, essay scores had large main effects on essay length (η_p^2 = 0.35), lexical diversity (η_p^2 = 0.33), and syntactic variety (η_p^2 = 0.25) and medium main effects on age of acquisition (η_p^2 = 0.08) and imageability (η_p^2 = 0.05; Cohen, 1988). In contrast, task types had small main effects on the five linguistic features (essay length: η_p^2 = 0.02; lexical diversity: η_p^2 = 0.02; age of acquisition: η_p^2 = 0.01; imageability: η_p^2 = 0.01; syntactic variety: η_p^2 = 0.03). These findings indicate that variations in students' linguistic complexity were more pronounced across essay scores than between task types.

Furthermore, significant interactions between essay score and task type were observed for essay length (η_p^2 = 0.07), lexical diversity (η_p^2 = 0.05), and syntactic variety (η_p^2 = 0.03), albeit with small to medium effect sizes. No significant interactions were found for age of acquisition

and imageability. This finding suggested that variations across essay scores in essay length, lexical diversity, and syntactic variety differed significantly between descriptive and email tasks.

Table 4. Results of Large Language Models Predicting Linguistic Features Across Essay Scores

Effect	Estimate (*SE*)				
	Model 1 Essay length [a]	Model 2 Lexical diversity	Model 3 Age of acquisition	Model 4 Imageability	Model 5 Syntactic variety
Fixed effect					
(Intercept)	-1.06 (0.12)*	-1.37 (0.12)*	-0.65 (0.12)*	0.53 (0.12)*	-1.12 (0.13)*
Score2	0.67 (0.12)*	0.82 (0.13)*	0.36 (0.13)*	-0.33 (0.13)*	0.50 (0.14)*
Score3	1.68 (0.13)*	1.73 (0.14)*	0.80 (0.13)*	-0.58 (0.13)*	1.38 (0.15)*
Score4	2.10 (0.15)*	2.25 (0.17)*	0.86 (0.14)*	-0.75 (0.14)*	1.77 (0.18)*
Email	0.29 (0.15)	0.71 (0.17)*	0.17 (0.07)*	-0.15 (0.07)*	0.54 (0.19)*
Score2: Email	-0.24 (0.18)	-0.32 (0.20)			0.05 (0.21)
Score3: Email	-0.72 (0.18)*	-0.66 (0.20)*			-0.36 (0.22)
Score4: Email	-1.01 (0.20)*	-1.04 (0.22)*			-0.61 (0.24)*
Random effect					
Var: Student (Intercept)	0.19	0.09	0.03	0.05	0.07
Var: Residual	0.39	0.50	0.86	0.88	0.59

Note. Score2 = essay score of two; Score3 = essay score of three; Score4 = essay score of four; Var = variance; *SE* = standard error. Number of individual students = 114; total *N* = 684.
[a] Values of essay length are after square-root-transformation.
* $p < .05$

Given the significant interaction effects for essay length, lexical diversity, and syntactic variety in the final models, we performed Tukey's HSD post-hoc pairwise comparisons between essay scores by task (see Table 5). The results showed that all five linguistic features significantly increased from Score 1 to Score 2 and from Score 2 to Score 3 for both tasks. However, the increases between Scores 3 and 4 were generally statistically nonsignificant—with the exception of essay length in the descriptive tasks.

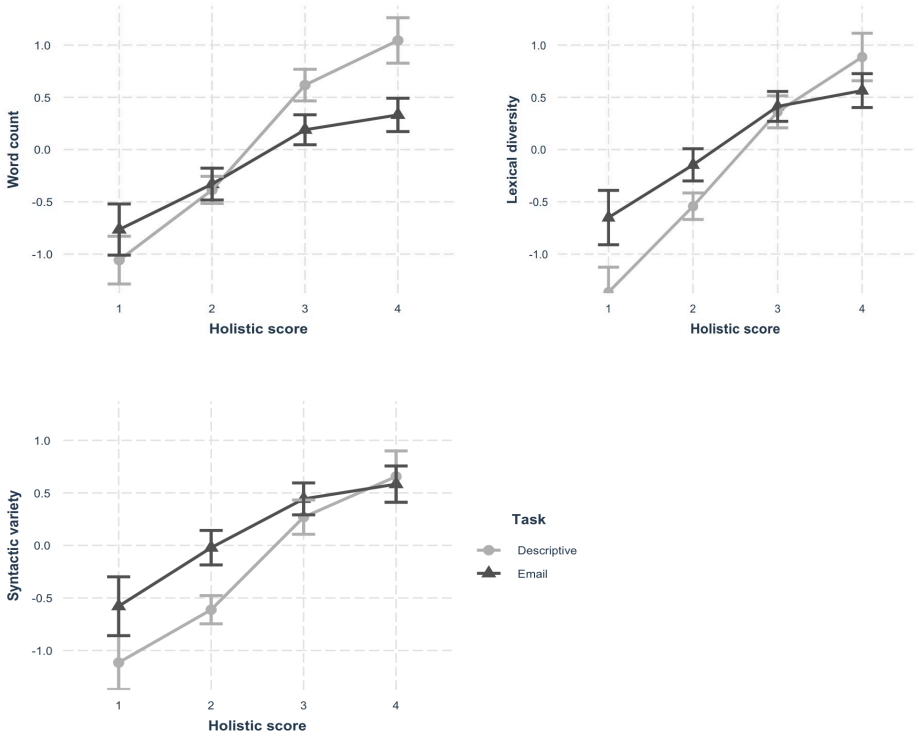
Table 5. Post-Hoc Pairwise Comparisons by Essay Score and Task Type

Contrast score	Estimate (*SE*)					
	Model 1: Essay length [a]		Model 2: Lexical diversity		Model 5: Syntactic variety	
	Descriptive	Email	Descriptive	Email	Descriptive	Email
1 → 2	0.67 (0.12)*	0.44 (0.14)*	0.82 (0.13)*	0.50 (0.15)*	0.50 (0.14)*	0.56 (0.16)*
2 → 3	1.00 (0.09)*	0.52 (0.09)*	0.90 (0.10)*	0.56 (0.10)*	0.88 (0.10)*	0.46 (0.11)*
3 → 4	0.43 (0.12)*	0.14 (0.10)	0.52 (0.14)*	0.15 (0.11)	0.39 (0.15)	0.14 (0.11)
1 → 3	1.68 (0.13)*	0.96 (0.14)*	1.73 (0.14)*	1.06 (0.15)*	1.38 (0.15)*	1.02 (0.16)*
2 → 4	1.43 (0.12)*	0.66 (0.10)*	1.43 (0.13)*	0.71 (0.11)*	1.27 (0.14)*	0.60 (0.12)*
1 → 4	2.10 (0.16)*	1.10 (0.14)	2.25 (0.17)*	1.22 (0.15)*	1.77 (0.18)*	1.16 (0.17)*

Note. Number of individual students = 114; total *N* = 684. *SE* = standard error.
[a] Values of essay length are after square-root-transformation.
* $p < .05$

Figure 3 displays how these linguistic features varied across holistic scores for the two tasks. As shown in the figure, the increases in essay length, lexical diversity, and syntactic variety were more pronounced between Scores 1 and 2 and between Scores 2 and 3 than between Scores 3 and 4 for both tasks. Furthermore, feature variations across holistic essay scores tended to be greater in the descriptive task compared to the email task.

Figure 3. Essay Length, Lexical Diversity, and Syntactic Variety by Essay Score, Task Type



Interestingly, the differences in lexical diversity and syntactic variety between task types diminished progressively from Score 1 to Score 4. However, differences in essay length across task types became larger at higher score levels (Scores 3 and 4). Notably, at the score level of 3, descriptive essays began to surpass email essays in length, and those receiving a score of 4 exhibited higher lexical diversity and slightly greater syntactic variety.

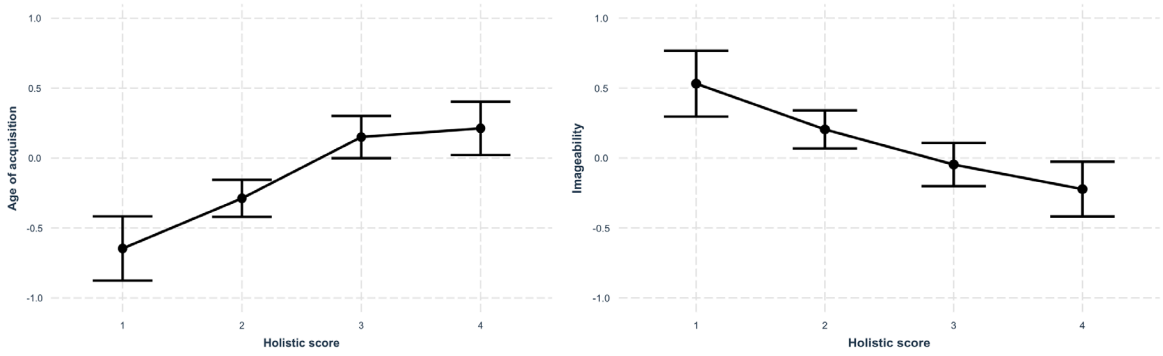
In contrast, the final models for age of acquisition and imageability did not yield a significant interaction effect. Consequently, we performed Tukey's HSD post-hoc pairwise comparisons between holistic scores (see Table 6) and plotted these features across scores by combining the two tasks (see Figure 4). Age of acquisition increased significantly from lower to higher score levels whereas imageability exhibited the opposite trend (see Table 6). This finding indicated that higher scoring essays tended to include words acquired later in life and less imageable or abstract words. Consistent with the patterns observed for the other three features, neither age of acquisition nor imageability showed significant variations from Score 3 to Score 4. Also, the magnitude of these differences diminished at the higher score levels (see Figure 4).

Table 6. Post-Hoc Pairwise Comparisons for Age of Acquisition, Imageability by Essay Score

Contrast score	Estimate (*SE*)	
	Model 3: Age of acquisition	Model 4: Imageability
1 → 2	0.36 (0.13)*	-0.33 (0.13)
2 → 3	0.44 (0.09)*	-0.25 (0.09)*
3 → 4	0.06 (0.11)	-0.18 (0.11)
1 → 3	0.80 (0.13)*	-0.58 (0.13)*
2 → 4	0.50 (0.10)*	-0.43 (0.11)*
1 → 4	0.86 (0.14)*	-0.75 (0.15)*

Note. *SE* = standard error. Number of individual students = 114; total *N* = 684.
* $p < .05$

Figure 4. Age of Acquisition and Imageability Across Essay Scores



In summary, the written responses with higher essay scores demonstrated more complex linguistic features, including longer length, greater lexical diversity, later age of acquisition, lower imageability, and increased syntactic variety. Interestingly, the differences between higher scoring essays were smaller than those observed between lower scoring ones. The email task elicited more linguistically complex responses compared to the descriptive task. No statistically significant effects of education level were found on the linguistic features examined. Task type appeared to mediate the relationship between linguistic features and holistic essay scores. Finally, the variations in linguistic features across essay scores were greater in the descriptive responses than in email responses.

RQ2. How Do the Linguistic Features in Young EAL Learners' Descriptive and Email Essays Change Over the Course of 1 Year?

Table 7 presents the descriptive statistics of the five linguistic features assessed across the two writing tasks during the three test administrations. The results showed that responses from later administrations differed from those at earlier administrations in length, lexical diversity, average age of acquisition, imageability, and sentence variety. These findings suggested that young EAL learners' writing exhibited noticeable linguistic changes over a 1-year period.

Table 7. Descriptive Statistics of Linguistic Features Across Three Administrations

Time	*n*	Measurement				
		Essay length	Lexical diversity	Age of acquisition	Imageability	Syntactic variety
0	228	65.05 (45.40)	3.62 (0.84)	2.41 (0.72)	3.55 (0.25)	1.57 (0.66)
1	228	97.37 (50.81)	4.01 (0.69)	2.67 (0.61)	3.41 (0.16)	2.03 (0.54)
2	228	110.82 (64.64)	4.12 (0.79)	2.62 (0.56)	3.46 (0.18)	2.19 (0.57)

Note. Number of individual students = 114; total *N* = 684.

Using the same approach, LMMs were selected for each linguistic feature to examine changes over time and differences between task types and education levels. All final models, except the one predicting essay length, met the LMM model assumptions. To address the violation of homoscedasticity, the essay length variable was square-root-transformed and standardized, which solved the issue.

Table 8 presents the LMM results for each linguistic feature. The reference levels were Time 0 (i.e., the first administration of the writing tasks), the descriptive task, and primary school. The results indicated that time had a statistically significant effect on all five linguistic

features. The effect of time on essay length included a random slope, suggesting substantial variation across individual students in how their essay length changed over time. All features except essay length were significantly affected by task type, consistent with findings from models using holistic essay scores as reported in the previous section. Among the five linguistic features, only age of acquisition was significantly affected by education level ($\eta_p^2 = 0.02$). As shown in Table 8, secondary school students tended to use words acquired later in life compared to primary school students, aligning with expectations.

Despite differences due to task type and education level, as time progressed students' essays generally tended to be longer with a broader range of vocabulary, more words acquired later in life, less imageable words, and greater sentence variety. These findings again reflected the meaningful linguistic changes in students' written responses over a 1-year period.

In effect sizes, time had the largest main effect on syntactic variety ($\eta_p^2 = 0.25$), followed by essay length ($\eta_p^2 = 0.22$), lexical diversity ($\eta_p^2 = 0.13$), imageability ($\eta_p^2 = 0.10$), and age of acquisition ($\eta_p^2 = 0.03$). In contrast, task type and education level exerted relatively small effects. Among task type effects, syntactic variety was most strongly impacted ($\eta_p^2 = 0.12$), followed by lexical diversity ($\eta_p^2 = 0.07$), imageability ($\eta_p^2 = 0.02$), and age of acquisition ($\eta_p^2 = 0.02$). These findings suggested that temporal changes in students' linguistic complexity were more pronounced than variations attributable to task type or education level.

Furthermore, there were statistically significant interaction effects between time and task type on age of acquisition ($\eta_p^2 = 0.01$), imageability ($\eta_p^2 = 0.07$), and syntactic variety ($\eta_p^2 = 0.01$), with small to medium effect sizes. These results indicated that the changes in these linguistic features across the three testing periods differed between the two writing tasks. To further explore these interactions, we conducted Tukey's HSD post-hoc pairwise comparisons by time and task type (see Table 9). Significant changes in all three features were observed from Time 0 to Time 1 and for both tasks. However, from Time 1 to Time 2, significant linguistic changes were limited to imageability in the email task and syntactic variety in the descriptive task. From Time 0 to Time 2, the descriptive task responses showed significant changes in all three linguistic features, whereas the email task responses displayed significant changes only in syntactic variety. Interestingly, no significant changes in age of acquisition were observed in the email task responses over time.

Table 8. Results of the LMMs Predicting Linguistic Features Over Time

Effect	Estimate (*SE*)				
	Model 1 Essay length [a]	Model 2 Lexical diversity	Model 3 Age of acquisition	Model 4 Imageability	Model 5 Syntactic variety
Fixed effect					
(Intercept)	-0.50 (0.07)*	-0.56 (0.08)*	-0.76 (0.12)*	0.77 (0.09)*	-0.90 (0.08)*
Time1	0.67 (0.06)*	0.49 (0.07)*	0.62 (0.12)*	-1.00 (0.11)*	0.84 (0.09)*
Time2	0.85 (0.08)*	0.62 (0.07)*	0.38 (0.12)*	-0.97 (0.11)*	1.12 (0.09)*
Email		0.37 (0.06)*	0.46 (0.12)*	-0.81 (0.11)*	0.68 (0.09)*
Secondary school			0.38 (0.10)*		
Time1: Email			-0.46 (0.17)*	0.65 (0.16)*	-0.25 (0.13)
Time2: Email			-0.12 (0.17)	1.08 (0.16)*	-0.31 (0.13)*
Random effects					
Var: Student (Intercept)	0.50	0.34	0.07	0.11	0.27
Var: Student Time1	0.21				
Var: Student Time2	0.50				
Cov: Student (Intercept) Time1	-0.12				
Cov: Student (Intercept) Time2	-0.09				
Cov: Student Time1 Time2	0.25				
Var: Residual	0.27	0.55	0.85	0.75	0.50

Note. Number of individual students = 114; total *N* = 684. Time1 = Time 1; Time2 = Time 2; Var = variance; Cov = covariance; *SE* = standard error.

[a] Values of essay length are after square-root-transformation.

* $p < .05$

Table 9. Post-Hoc Pairwise Comparisons for Age of Acquisition, Imageability, and Syntactic Variety by Time and Task Type

Contrast time	Estimate (*SE*)					
	Model 3: Age of acquisition		Model 4: Imageability		Model 5: Syntactic variety	
	Descriptive	Email	Descriptive	Email	Descriptive	Email
0 → 1	0.62 (0.12)*	0.16 (0.12)	-1.00 (0.12)*	-0.35 (0.12)*	0.84 (0.09)*	0.59 (0.09)*
1 → 2	-0.25 (0.12)	0.09 (0.12)	0.04 (0.12)	0.47 (0.12)*	0.29 (0.09)*	0.22 (0.09)
0 → 2	0.38 (0.12)*	0.25 (0.12)	-0.97 (0.12)*	0.12 (0.12)	1.12 (0.09)*	0.81 (0.09)*

Note. Number of individual students = 114; total *N* = 684. *SE* = standard error.
* $p < .05$

Figure 5 illustrates changes over time in three linguistic features by task: syntactic variety, age of acquisition, and imageability. All three features were significantly influenced by interactions between time and task type. The figure reveals a consistent increase in syntactic variety scores from Time 0 to Time 2 for both tasks, indicating that students' descriptive and email essays exhibited an increasingly wider range of sentence structures over the 1-year period. However, the rate of increase slowed from Time 1 to Time 2 compared to the earlier period from Time 0 to Time 1. In contrast, nonlinear change patterns were observed for age of acquisition and imageability. Although students' essays appeared to contain more sophisticated words from Time 0 to Time 1 for both tasks, this trend did not persist between Time 1 and Time 2.

Since there was no statistically significant interaction between time and task type for essay length and lexical diversity, we conducted Tukey's HSD post-hoc pairwise comparisons across time for these two features combining data from both task types (see Table 10 and Figure 6). Overall, essay length and lexical diversity increased steadily over time. Notably, as illustrated in Figure 6, the rates of change, as indicated by the slopes, were faster between Time 0 and Time 1 for both features, compared to the slower progress observed between Time 1 and Time 2. This mirrors the pattern observed for syntactic variety in Figure 5, regardless of task type.

Figure 5. Changes of Age of Acquisition, Imageability, and Syntactic Variety by Task Type

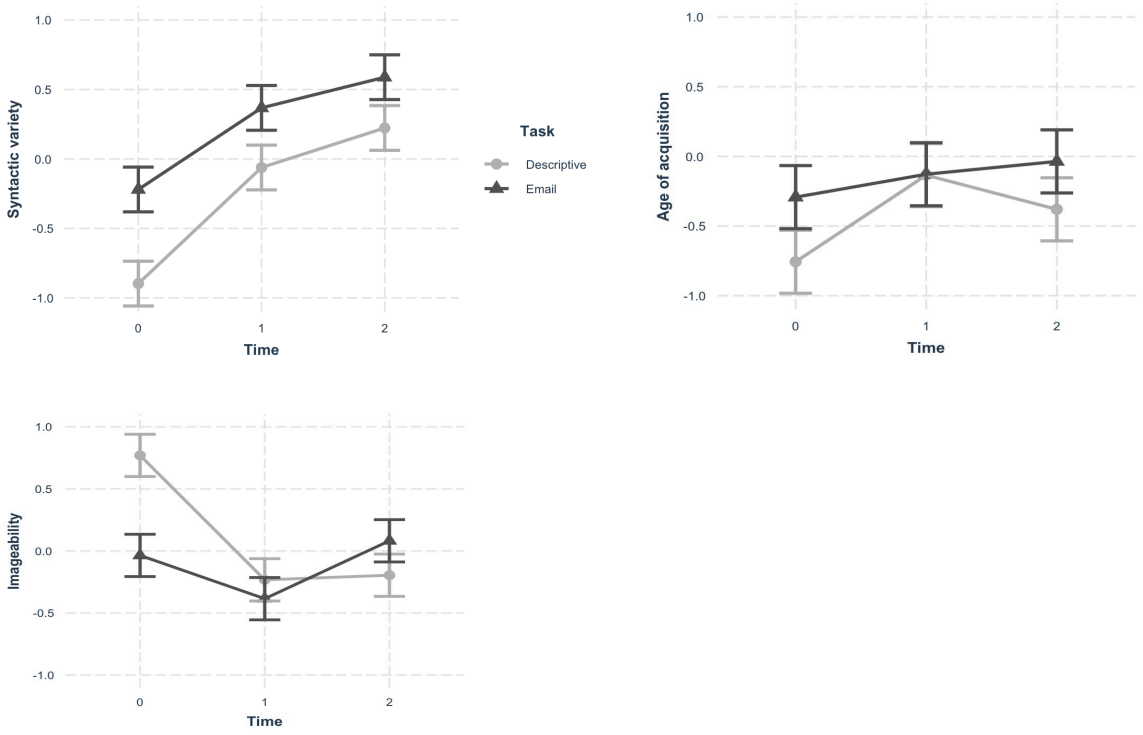
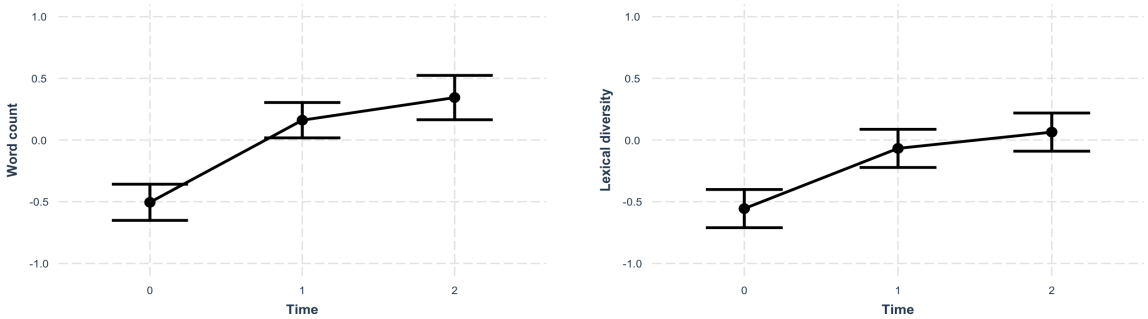


Table 10. Post-Hoc Pairwise Comparisons for Essay Length and Lexical Diversity by Time

Contrast time	Estimate (*SE*)	
	Model 1: Essay length [a]	Model 2: Lexical diversity
0 → 1	0.67 (0.07)*	0.48 (0.07)*
1 → 2	0.18 (0.07)*	0.13 (0.07)
0 → 2	0.85 (0.07)*	0.62 (0.07)*

Note. Number of individual students = 114; total *N* = 684. *SE* = standard error.
[a] Values of essay length are after square-root-transformation.
* $p < .05$

Figure 6. Changes of Essay Length and Lexical Diversity for Both Tasks



Taken together, over the 1-year period, the students' written texts became longer and included more advanced vocabulary and a wider range of sentence structures. The magnitude of differences in linguistic features was larger between Time 0 and Time 1 than between Time 1 and Time 2. We speculated that this slower change may coincide with the extended summer break between Time 1 and Time 2 when students did not receive formal language instruction in the classroom.

In terms of task effects, we observed that the email task generally elicited more complex language than the descriptive task, including greater lexical diversity, the use of more concrete words, words acquired later in life, and a wider variety of syntactic structures. However, the descriptive task might offer a greater scope for responses at higher score levels to demonstrate length and linguistic complexity, as higher scoring descriptive essays tended to be longer with more diverse vocabulary and more varied sentence structures compared to the email responses. In addition, older students, particularly those in secondary schools, were more likely than primary school students to use words acquired later in life, providing empirical support for the age of acquisition measure. Task type also influenced linguistic changes over time with the descriptive responses exhibiting more variations compared to the email responses.

Discussion

In this study, we examined linguistic features exhibited in young EAL learners' written texts across score levels and task types (RQ1) and the changes in linguistic features across three test administrations over a 1-year period (RQ2). The following sections summarize the key findings in relation to each research question.

RQ1. Linguistic Features Across Score Levels and Task Types

To address the first research question, we investigated how a selected set of lexical and syntactic features varied across holistic task scores and task types. Our findings showed clear and consistent score-related differences in all five linguistic features examined. Higher scoring responses were generally longer, demonstrated greater lexical diversity, contained words acquired later in life, involved less imageable words, and exhibited greater syntactic variety. These observed score-related differences align with prior cross-sectional research on both adult and young L2 writers, which has shown that higher quality texts tend to exhibit more complex

lexical and syntactic characteristics (e.g., Bulté & Housen, 2014; Kyle & Crossley, 2017; Maamuujav, 2021; Wolf et al., 2018).

Task-based language teaching indicates that task design variables, such as task complexity, may influence linguistic complexity of language production (Ellis et al., 2019; Yoon, 2021). In this study, we found that task type played a role in shaping the observed score-related linguistic variations. The email task elicited responses with greater lexical diversity, more words acquired later in life, lower imageability, and greater syntactic complexity than the descriptive task, particularly at lower score levels. On the other hand, the descriptive task appeared to allow more differentiation at higher score levels, particularly in essay length and syntactic variety. However, differences between task types diminished at higher score levels. Taken together, our findings imply that as writing quality increases, young EAL learners' written texts may become more similar in their linguistic characteristics across task types. Additionally, linguistic features in children's written texts can distinguish stronger and weaker writing performances both within and across task types.

RQ2. Linguistic Changes Across Administrations

The second research question focused on how the same set of linguistic features changed across three test administrations over a 1-year period. The results showed that, overall, students' written responses at later administrations differed from those at earlier administrations in length, lexical diversity, average age of acquisition, and syntactic variety. The rate and magnitude of change, however, differed by linguistic features. Overall, students' essay lengths, lexical diversity, and syntactic variety showed steady increase over time. In contrast, age of acquisition and imageability displayed more nuanced, nonlinear patterns, with significant changes taking place primarily between the first and second administrations; weaker or nonsignificant changes took place between the second and third administrations. These patterns suggest that different linguistic features may change or develop at different rates and may be differentially sensitive to children's formal instructional exposure or other broader contextual factors.

Across the five linguistic features, it is worth pointing out that the most pronounced changes occurred between the first and second administrations and changes between the second and third administrations were slower and more variable. This pattern coincided with an extended summer break between the latter two timepoints, during which students received limited formal classroom instruction. This finding is consistent with previous research that shows

children's productive use of language does not necessarily progress in a linear fashion and can be influenced by fluctuations in instructional input (e.g., Levitzky-Aviad & Laufer, 2013). Our finding also echoes the dynamic systems theory that emphasizes (a) variability and nonlinearity in SLA and (b) that different linguistic subsystems may interact and reorganize over time rather than develop uniformly (Larsen-Freeman, 2006; Verspoor et al., 2008).

Additionally, the results indicated that task type mediated the longitudinal linguistic change patterns. Although both descriptive and email tasks showed overall increases in syntactic complexity, changes in age of acquisition and imageability differed across tasks and there was more consistent change in the descriptive task. This finding suggests that task type may influence not only how young EAL learners use language but also how linguistic features vary over time.

Pedagogical Implications

From a pedagogical perspective, findings of the study have several implications for supporting young EAL writers. First, our findings indicate that exposure to varied task types can support richer linguistic performance in young students' writing. Different writing tasks appear to elicit distinct linguistic features, highlighting the value of incorporating a range of writing task types or genres in classroom writing instruction. For example, teachers might consider using correspondence writing that involves written interaction (e.g., email, online discussion, text messages) to foster different aspects of linguistic development. Additionally, descriptive tasks may be utilized to support narrative development and facilitate the use of different sentence structures, and email tasks can be used to teach audience awareness or real-life written communication. Teachers may also scaffold writing tasks by introducing task-specific vocabulary and sentence structures that meet task requirements and support the development of students' writing skills. Furthermore, the variations in linguistic features across score levels highlight the need for differentiated instruction. Teachers may use classroom-based writing assessment tasks to identify individual students' strengths and weaknesses at different score levels and tailor their instruction to support specific linguistic or developmental needs, such as vocabulary development or syntactic control.

The study findings also provide empirical support for age-related differences in lexical choice. As shown in our data, compared to primary school students, secondary school students were more likely to use words acquired later in life, even when responding to the same writing tasks. This pattern is consistent with prior psycholinguistic research that shows that compared to

younger children, older students have better access to and can deploy a lexicon that includes more abstract or less frequent words (Bird et al., 2001; Kuperman et al., 2012). This finding also demonstrates that age of acquisition can be used as a meaningful measure of lexical sophistication in young EAL learners' written responses.

Conclusion

This study examined linguistic features in young EAL learners' descriptive and email writing by integrating automated linguistic analyses with human ratings of writing samples in a classroom-based longitudinal design. Findings of the study provide insight into how linguistic characteristics in children's descriptive and email writing are associated with essay scores and how these characteristics change across three administrations over a 1-year period. With respect to score-related variations, higher scoring responses consistently differed from lower scoring responses across the linguistic dimensions examined. Regarding changes across test administrations, findings showed that students' written responses produced at later administrations were generally longer and exhibited greater lexical and syntactic variety than those produced at earlier administrations. Results also revealed that the magnitude and rate of change differed by linguistic feature, with syntactic variety showing the most consistent increase and lexical sophistication measures displaying more nuanced patterns. Changes were more pronounced between the first and second administrations and slower between the second and third.

Several limitations of the study need to be acknowledged. First, we analyzed young EAL students' written texts collected from three European learning contexts. The study samples may limit the generalizability of the findings to other cultural and educational contexts. In addition, the small sample size did not allow for cross-linguistic comparison and may also affect the robustness of the findings from the analyses with LMMs. Future research should include more diverse and larger young EAL student populations to explore cross-cultural, cross-linguistic differences in EAL learners' linguistic changes and development.

Second, we only included two task types and five NLP-based linguistic features in our analyses. Expanding the scope of the analyses to include more task types and linguistic phenomena such as cohesion and coherence will provide a more comprehensive understanding of young EAL learners' linguistic characteristics and writing performances. Finally, the observed nonlinear change patterns in some linguistic features warrant further investigation. Longitudinal

studies spanning longer periods of time can reveal more detailed linguistic change trajectories. Research should also explore the impact of instructional interventions on linguistic change and writing development over a longer period of time.

It is also important to note that future research should isolate linguistic change from prompt-related variability. Researchers may adopt a counterbalanced design such that each prompt is administered at multiple waves to different subgroups of students. This approach would allow researchers to more directly examine whether observed changes are attributable to time, task characteristics, or their interaction. Because the present study did not employ such a design, observed changes across time should be interpreted as differences in response characteristics across administrations rather than definitive evidence of individual writing development.

Despite the limitations, the study provides insight into young EAL learners' written responses, showing patterns of linguistic variations across score levels, task types, and test administrations. The findings highlight the importance of employing varied writing tasks and scaffolded linguistic instruction in teaching young EAL writers and contribute meaningfully to our understanding of early second language writing development.[2]

References

Attali, Y., & Burstein, J. (2006). Automated scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 3–30. https://ejournals.bc.edu/index.php/jtla/article/view/1650/1492

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL® 2000 spoken and written academic language corpus* (ETS Research Memorandum RM-04-03). ETS. http://www.ets.org/Media/Research/pdf/RM-04-03.pdf

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers, 33*, 73–79. https://doi.org/10.3758/BF03195349

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42–65. http://dx.doi.org/10.1016/j.jslw.2014.09.005

Bulté, B., & Housen, A. (2017). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics, 28*(1), 147–164. https://doi.org/10.1111/ijal.12196

Butler, Y. G. (2025). *Children's additional language learning in instructional settings: Implications for teaching and future research*. Multilingual Matters. https://doi.org/10.21832/butler3788

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S.A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282–311. https://doi.org/10.1177/0741088311410188

De Wilde, V. (2023). Lexical characteristics of young L2 English learners' narrative writing at the start of formal instruction. *Journal of Second Language Writing, 59*, Article 100960. https://doi.org/10.1016/j.jslw.2022.100960

Ellis, R., Skehan, P., Li, S., Shintani, N., & Lambert, C. (2019). *Task-based language teaching: Theory and practice*. Cambridge University Press. https://doi.org/10.1017/9781108643689

Erasmus. (2022, September 14). *Innovating & Building Understanding to Improve Language Development with Information Technology Erasmus project.* https://www.erasmusly.com/innovating-building-understanding-to-improve-language-development-with-information-technology-erasmus-project.html

ETS. (2023). *TOEFL Primary® Writing scoring guide*. https://www.ets.org/pdfs/toefl/toefl-primary-writing-scoring-guide.pdf

ETS. (2025). *TOEFL Junior® Writing scoring guide*. https://www.ets.org/pdfs/toefl/toefl-junior-writing-scoring-guide.pdf

ETS. (2026). *Prepare for the TOEFL Junior® tests*. https://www.ets.org/toefl/junior/prepare.html

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

Gené-Gil, M., Juan-Garau, M., & Salazar-Noguera, J. (2015). Development of EFL writing over three years in secondary education: CLIL and non-CLIL settings. *The Language Learning Journal, 43*(3), 286–303. http://dx.doi.org/10.1080/09571736.2015.1053278

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Houghton Mifflin.

Hsieh, C.-N. (2024). The role of task types and reading proficiency on young English as a foreign language learners' writing performances. *TESOL Quarterly*, *58*(2), 978–990. https://doi.org/10.1002/tesq.3286

Jiang, J., Bi, P., & Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing, 46*, Article 100666. https://doi.org/10.1016/j.jslw.2019.100666

Kim, M. (2021). Exploring longitudinal changes in lexical and syntactic features in beginning-level EFL learner writing. *System, 103*, Article 102680. https://doi.org/10.1016/j.system.2021.102680

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

Kyle, K., & Crossley, S. (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

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal, 102*(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition, 43*(4), 781–812. https://doi.org/10.1017/S0272263120000546

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590–619. https://doi.org/10.1093/applin/aml029

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. In C. Bardel, C. Lindqvist, & B. Laufer, B. (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 127–148). EuroSLA-the European Second Language Association. https://www.eurosla.org/monographs/EM02/EM02tot.pdf

Loder Buechel, L. (2015). Young learner writing performance in Swiss elementary schools—Which teacher variables matter? *RELC Journal*, *46*(3), 275–292. https://doi.org/10.1177/0033688215598437

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Maamuujav, U. (2021). Examining lexical features and academic vocabulary use in adolescent L2 students' text-based analytical essays. *Assessing Writing*, *49*, Article 100540. https://doi.org/10.1016/j.asw.2021.100540

Maamuujav, U., Olson, C. B., & Chung, H. (2021). Syntactic and lexical features of adolescent L2 students' academic writing. *Journal of Second Language Writing, 53*, Article 100822. https://doi.org/10.1016/j.jslw.2021.100822

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Multilingual Matters.

Martínez, A. C. L. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing, 35*, 1–11. https://doi.org/10.1016/j.asw.2017.11.002

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing, 29*, 3–15. http://dx.doi.org/10.1016/j.jslw.2015.06.004

Michel, M., Kormos, J., Brunfaut, T., & Ratajczak, M. (2019). The role of working memory in young second language learners' written performances. *Journal of Second Language Writing*, *45*, 31–45. https://doi.org/10.1016/j.jslw.2019.03.002

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing, 29*, 82–94. http://dx.doi.org/10.1016/j.jslw.2015.06.008

Patekar, J. (2021). A look into the practices and challenges of assessing young EFL learners' writing in Croatia. *Language Testing*, *38*(3), 456–479. https://doi.org/10.1177/0265532221990657

Pinter, A. (2006). *Teaching young language learners*. Oxford.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards* (Research Report Series No. RR-10-28). ETS. https://doi.org/10.1002/j.2333-8504.2010.tb02235.x

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd edition). Sage.

Sung, H., Wolf, M. K., Suhan, M., & Kyle, K. (2025). Lexical richness in young English learners' writing: A focus on opinion and listen-write task types. *Assessing Writing*, *66*, Article 100975. https://doi.org/10.1016/j.asw.2025.100975

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly, 47*(2), 420–430. https://doi.org/10.1002/tesq.91

Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal, 92*(2), 214–231. https://psycnet.apa.org/doi/10.1111/j.1540-4781.2008.00715.x

Vogt, K., Tsagari, D., Csépes, I., Green, A., & Sifakis, N. (2020). Linking learners' perspectives on language assessment practices to teachers' assessment literacy enhancement (TALE): Insights from four European countries. *Language Assessment Quarterly*, *17*(4), 410–433. https://doi.org/10.1080/15434303.2020.1776714

Wang, Z. (2022). Dynamic development of syntactic complexity in second language writing: A longitudinal case study of a young Chinese EFL learner. *Frontiers in Psychology, 13*, Article 974481. https://doi.org/10.3389/fpsyg.2022.974481

Wolf, M. K., Oh, S., Wang, Y., & Tsutagawa, F. S. (2018). Young adolescent EFL students' writing skill development: Insights from assessment data. *Language Assessment Quarterly*, *15*(4), 311–329. https://doi.org/10.1080/15434303.2018.1531868

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67. http://dx.doi.org/10.1016/j.jslw.2015.02.002

Yoon, H.-J. (2021). Challenging the connection between task perceptions and language use in L2 writing: Genre, cognitive task complexity, and linguistic complexity. *Journal of Second Language Writing, 54,* Article 100857. https://doi.org/10.1016/j.jslw.2021.100857

Yoon, H.-J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly, 51*(2), 275–301. https://doi.org/10.1002/tesq.296

Young, D. S. (2017). *Handbook of regression methods*. CRC Press. https://doi.org/10.1201/9781315154701

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System, 56*, 40–53. https://doi.org/10.1016/j.system.2015.11.007

Notes

[1] The TOEFL Junior Writing test was launched after the project completed data collection.
[2] We would like to acknowledge the use of an AI tool to enhance readability of the manuscript.

Suggested Citation

Hsieh, C.-N., & Cai, H. (2026). *Linguistic features in young English-as-an-additional-language learners' writing: A classroom-based longitudinal study* (TOEFL Research Report Series No. TOEFL-RR-110). ETS. https://doi.org/10.64634/0t8dx119

Action Editor: Jonathan Schmidgall

Reviewers: Veronika Tempe-Laughlin and Saerhim Oh

E-RATER, ETS, the ETS logo, TEXTEVALUATOR, TOEFL, TOEFL IBT, TOEFL JUNIOR, and TOEFL PRIMARY are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database.

www.ets.org

Copyright © 2026 by ETS. The Eight-Point logo is a trademark of ETS. All other trademarks are property of their respective owners.