Korean Journal of English Language and Linguistics, **Vol 25, October 2025, pp. 1378-1394**
DOI: 10.15738/kjell.25..202510.1378

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Exploring Vocabulary Patterns in ESL and EFL Learner Corpora: Implications for Language Teaching*

Seulgi Choi · Sunyoung Park · Jin-young Tak (Sejong University)

Received: July 18, 2025
Revised: August 22, 2025
Accepted: October 16, 2025

Choi, Seulgi (First author)
Master's Student, Department of Artificial Intelligence and Language Engineering
Sejong University
Email: csg0726@sju.ac.kr

Park, Sunyoung (Co-author)
Visiting Professor, Daeyang Humanity College
Sejong University
Email: sunpark@sejong.ac.kr

Tak, Jin-young (Corresponding author)
Professor, English Data Convergence Major
Sejong University
Email: jytak@sejong.ac.kr

* Seulgi Choi and Sunyoung Park contributed equally to the conception and design of the study and share first authorship.

ABSTRACT

Choi, Seulgi, Sunyoung Park and Jin-young Tak. 2025. Exploring vocabulary patterns in ESL and EFL learner corpora: Implications for language teaching. *Korean Journal of English Language and Linguistics 25*, 1378-1394.

Vocabulary knowledge is a crucial determinant of writing proficiency and overall communicative competence for second language learners. While previous research examined lexical richness in academic writing, few studies have directly compared English as a Second Language (ESL) and English as a Foreign Language (EFL) learners. This study attempts to fill that gap by analyzing vocabulary use in written texts from EFL learners and ESL learners, focusing on how different learning environments influence lexical development. Using the Linguistic Feature Toolkit (LFTK), an open-source tool with over 220 linguistically defined features, and the International Corpus Network of Asian Learners of English (ICNALE), the study systematically examines lexical features across groups. The results of the current study found interesting patterns between ESL and EFL learners. ESL learners, immersed in English speaking contexts, demonstrate greater lexical flexibility and diversity in content word usage that closely aligns with native speaker patterns. In contrast, EFL learners exhibited more native-like patterns in functional word usages. This phenomenon can be interpreted as that ESL learners develop greater diversity in content word usage as a result of immersive exposure, whereas EFL learners, through form-focused instruction, developed functional word usage patterns more similar to native speakers. These results have important pedagogical implications: AI-based vocabulary learning tools can simulate immersive exposure for EFL learners, promoting lexical diversity and motivation, while ESL learners may benefit from explicit instruction targeting functional words to improve grammatical precision. Ultimately, the study emphasizes the importance of tailoring instruction to learners' environments to support balanced language development and effective writing outcomes.

KEYWORDS

vocabulary development, lexical diversity, EFL learners, ESL learners, writing proficiency, pedagogical implications

1. Introduction

The significance of vocabulary knowledge in enhancing writing competence for second language learners has been a key topic of research for decades (Bestgen 2017, Engber 1995, Ha 2019, Laufer and Nation 1995, Nation and Nation 2001, Schmitt 2008, among others). Studies consistently show that higher quality writing is often characterized by increased lexical complexity, which plays a crucial role in enhancing the depth and precision of expression (Bulté and Housen 2014, Lu 2010, O'Leary and Steinkrauss 2022, Ortega 2003, Zhang and Ouyang 2023). Vocabulary knowledge has been recognized as essential for successful engagement in all four fundamental language skills. Schmitt (2008) emphasizes that it is a core component for English learners, while Azodi et al. (2014) similarly note that a learner's vocabulary proficiency often reflects their overall fluency in listening, speaking, reading, and writing. González-López and López-López (2015) further highlight that vocabulary serves as a key indicator of effective academic writing and can be used to assess the quality of written texts. Additionally, Djiwandono (2016) points out that employing a diverse vocabulary in writing helps prevent a "monotonous and tedious tone" demonstrating that writers possess a robust lexical repertoire and familiarity with various English texts. Laufer and Nation (1995) also argue that the appropriate and effective use of extensive vocabulary positively impacts readers' perceptions.

In the context of English language education, these results highlight the critical role of vocabulary instruction. Teaching vocabulary not only improves students' writing but also enhances their overall language proficiency, enabling them to express ideas more clearly and creatively. Effective vocabulary instruction strengthens communication skills, boosts learners' confidence, and supports better academic performance. Thus, integrating vocabulary development into the curriculum is essential for fostering language skills and achieving long-term success in both academic and authentic communicative competence.

Although vocabulary instruction is important, it is often difficult to implement effectively without a comprehensive understanding of learners' vocabulary usage profiles. By examining how students use vocabulary in their writing, classroom teachers can tailor teaching strategies better to address specific needs and gaps. However, while previous research has focused on areas such as lexical richness, cross group comparisons, and longitudinal studies (Chang and Wang 2016, Johnson et al. 2012, Lu 2010, Norris and Ortega 2009), a notable gap remains in comparative studies between ESL (English as a Second Language) and EFL (English as a Foreign Language) learners regarding vocabulary acquisition.

The present study attempts to address this gap by comparing the vocabulary usage patterns in the written language of ESL and EFL learners. ESL learners, who are immersed in English-speaking environments, acquire language through frequent exposure to authentic input, whereas EFL learners develop their skills primarily in classroom-based contexts with limited access to such input. By analyzing their written essays, this study identifies distinct lexical patterns shaped by these differing learning conditions and examines their implications for language pedagogy. The current study aims to inform curriculum design, refines instructional approaches, and supports more targeted interventions to enhance vocabulary development in diverse learning environments.

2. Literature Review

One significant distinction in the field of English language acquisition and teaching is the differentiation between ESL and EFL contexts (Buschfeld and Kautzsch 2017, Loewen 2015, Ortega 2005, Shehadeh and Coombe 2012, among many others). This distinction is often framed in terms of learners' access to input and

opportunities for output. In ESL contexts, the target language is widely used within the surrounding community, providing learners with frequent and natural exposure to authentic language use. The idea that ESL contexts offer learners more frequent and naturalistic input exposure originates from early SLA research, particularly Krashen's Input Hypothesis (1985) and Long's Interaction Hypothesis (1981). More recent discussions of this perspective can be found in Loewen (2015) and Ortega (2005, 2015). In contrast, in EFL contexts, learners primarily encounter English in classroom settings, with limited opportunities for real-life communicative interaction. Consequently, ESL learners are typically assumed to benefit from richer input and more frequent opportunities for meaning negotiation, whereas EFL learners' exposure to the target language is often restricted to instructional materials and classroom discourse.

This difference in exposure has been widely linked to the development of learners' lexical knowledge. Numerous studies suggest that ESL learners tend to show greater lexical diversity and sophistication due to their immersion in naturalistic language environments (Vedder and Benigno 2016, Vold 2022). By comparison, EFL learners, whose vocabulary development is largely shaped by formal instruction, may exhibit a narrower lexical range but often demonstrate a higher reliance on academic or textbook-based vocabulary (Shehadeh and Coombe, 2012). These findings underscore the importance of considering the learning context when investigating L2 lexical development (Ortega 2005).

In recent decades, learner corpus research has emerged as a powerful methodological tool for examining lexical and syntactic features in L2 writing. Corpus-based studies provide empirical evidence of learners' actual language use and enable systematic comparisons across learner groups and proficiency levels. Measures of lexical complexity, including lexical diversity, lexical sophistication, and lexical density, have been widely applied to investigate how learners' lexical patterns vary according to context and proficiency (Kyle and Crossley 2015, Lu 2012). According to Ortega's (2003) overview of syntactic complexity research, linguistic development cannot be adequately represented by a single measure. Therefore, lexical complexity in L2 writing should also be approached as a multidimensional construct**.** Research has demonstrated that lexical complexity is not a unitary construct but a multidimensional phenomenon influenced by both linguistic and contextual factors.

One of the most significant contributions to learner corpus research in Asia is the International Corpus Network of Asian Learners of English (ICNALE), compiled by Ishikawa (2011, 2013). ICNALE provides a large and balanced collection of both spoken and written data from learners of diverse L1 backgrounds and proficiency levels across Asian EFL contexts. The corpus has been widely used to examine lexical and syntactic development in EFL learners' writing (Ishikawa 2011, 2013). Building on this foundation, Ishikawa (2024) conducted a corpus-based study examining the lexicogrammar of L2 English essays written by Asian college students. Using ICNALE data, the study analyzed lexical and grammatical features across learners from diverse L1 backgrounds through multidimensional analyses, cluster analysis, and statistical classification. The results indicated that learner essays exhibited distinct lexicogrammatical characteristics compared to native speakers, including lower levels of informational density, explicitness, and abstraction.

Ishikawa (2025) further examined the role of learners' L1 in shaping L2 writing by analyzing over one million words of ICNALE written data across 18 Asian learner groups. The study revealed that L1 does not always strongly determine learner output, and that other factors, such as proficiency, topic, and learning context, may exert greater influence. Together, these studies underscore the methodological and theoretical advances made in recent learner corpus research and highlight the importance of considering multiple variables beyond L1 when analyzing L2 writing.

Despite these advances, relatively few studies have employed corpus analysis to directly contrast ESL and EFL learners' writing, particularly with respect to lexical use. Therefore, the present study seeks to address this gap by

conducting a corpus-based comparison of ESL and EFL learners' written production. The current study examines lexical usage patterns across two learner groups using data from ICNALE. By systematically analyzing lexical usage patterns, this research attempts to contribute to a more comprehensive understanding of how language learning contexts shape learners' lexical knowledge. In doing so, it responds to recent calls in SLA research for integrated perspectives that consider both input-based theoretical models (Krashen 1985, Ortega 2003, 2005, 2015) and empirical corpus-based findings (Ishikawa 2024, 2025, Park 2023). In this context, the present study poses the following research questions:

1. To what extent does the context of language learning influence L2 acquisition?
2. Do discrepancies in written output exist between ESL and EFL learners?
3. How do ESL and EFL learners differ in their vocabulary use in L2 writing, particularly in terms of lexical choices, range, and frequency?

3. Method

The current study aims to examine linguistic differences in essay writing among Asian English learners, focusing on whether they are learning English as a Second Language (ESL) or as a Foreign Language (EFL). By analyzing the linguistic patterns demonstrated in their essays, the research seeks to inform the development of more effective teaching methodologies and support the design of tailored instructional approaches for each learner group.

3.1 Corpus Data

The International Corpus Network of Asian Learners of English (hereafter, ICNALE) is a large corpus comprising over 10,000 written and spoken data collected from undergraduate and graduate English learners across various Asian countries. Among the data provided by ICNALE, Written Essays (WE) v2.6, including learners' data from both ESL (Hong Kong, Pakistan, the Philippines, and Singapore) and EFL (China, Indonesia, Japan, Korea, Taiwan, and Thailand) countries, is investigated in this research. To ensure reliability in contrastive analysis, multiple variables, such as essay topics, writing time, and length were controlled. As for essay topics, participants chose one out of the following two standardized topics: (1) *whether part-time jobs are important for university students* and (2) *whether smoking should be completely banned in all restaurants*. The ICNALE corpus is particularly valuable for research comparing ESL and EFL learners, as it provides a large, standardized dataset from diverse Asian learner groups spanning both ESL and EFL contexts.

3.2 Dataset

Each dataset is categorized by proficiency level. Based on participants' English proficiency test scores (e.g., TOEIC, TOEFL) and an independent vocabulary test (Nation and Beglar, 2007), their proficiency levels were converted into four CEFR levels: A2, B1_1 (B1 low), B1_2 (B1 high), and B2+. The detailed structure of the data set used for analysis is presented in Tables 1 and 2.

Table 1. Profile of L2 learners in ESL and EFL

	Country	A2	B1-1	B1-2	B2+	Total
ESL	Hong Kong	2	60	104	34	200
	Pakistan	36	182	176	6	400
	Philippines	4	22	352	22	400
	Singapore	0	0	268	132	400
	Total	42	264	900	194	1400
EFL	China	100	464	210	26	800
	Indonesia	64	164	166	6	400
	Japan	308	358	98	36	800
	Korea	150	122	176	152	600
	Taiwan	58	174	122	46	400
	Thailand	238	358	200	4	800
	Total	918	1640	972	270	3800
	Total	960	1904	1872	464	5200

Table 1 summarizes the distribution of essay samples by country and proficiency level for learners in ESL and EFL contexts. The ESL group comprises 1,400 essays collected from four countries, while the EFL group includes 3,800 essays from six countries.

Table 2. Profile of NS

	Students	Teachers	Others	Total
NS (Native Speaker)	200	88	112	400

Table 2 presents the writer profiles and essay distribution of the NS (Native Speaker) reference group. The ICNALE corpus contains production data from native speakers (NS) of English, who completed the same tasks under identical conditions to those of non-native speakers (NNS), ensuring the reliability of NS–NNS comparisons. In collecting NS data, attention was paid to both geographical and demographic diversity. Consequently, three types of NS participants were included: (a) university students, (b) English teachers, and (c) adults from a range of professional backgrounds. Additionally, efforts were made to maintain a balanced representation of nationalities among NS participants: the United States (65% in the Speech Monologue Module / 57% in the Written Essays Module), the United Kingdom (17% / 14%), Australia (11% / 8.5%), Canada (7% / 6.5%), and others (>1% / 6.5%).

As shown in Table 1, the A2 level contained notably fewer samples in the ESL group (42 learners), creating a pronounced imbalance relative to the EFL group (918 learners). To avoid potential bias or distortion, this level was excluded from the analysis. While minor discrepancies in sample size persisted across the remaining levels, it was assumed that the group sizes were sufficiently robust for reliable statistical inference. In addition, given the potential violation of homogeneity of variance across groups, statistical methods accommodating unequal variances were employed. The analytical procedures are described in detail in Section 3.4.

3.3 Analysis on Linguistic Features

The analysis was conducted using Python, with each student essay serving as a unit of analysis for the extraction

of linguistic features. The features extracted from individual essays were then aggregated by groups which include ESL, EFL, and NS to construct a comprehensive dataset at the group level. Linguistic feature extraction was carried out using the Linguistic Feature ToolKit (LFTK), an open-source tool developed by Lee and Lee (2023) to support computational linguistic analysis. Built upon the spaCy library, LFTK systematically extracts and organizes over 220 handcrafted linguistic features that are widely used in natural language processing (NLP) research. Designed with extensibility and customizability in mind, the toolkit is applicable across a broad range of research contexts.

3.4 Statistical Testing

In this study, a series of statistical tests were conducted to examine whether there were significant differences in the mean values among the NS, ESL, and EFL groups. According to the Central Limit Theorem (CLT), even if the distribution of individual samples is non-normal, the distribution of the sample means approximates a normal distribution when the sample size is sufficiently large. Based on this assumption and the sufficiently large sample sizes in each group, mean-based statistical analyses were performed under the approximate assumption of normality ($\alpha = 0.05$).

Before analyzing the differences in group means, the homogeneity of variance was assessed. To this end, Levene's test was conducted using the SciPy library in Python. The results revealed that the assumption of homogeneity of variance was violated in approximately 67.73% of the features ($p < 0.05$). Based on these results, statistical methods that do not assume equal variances were adopted in the analysis.

Accordingly, Welch's ANOVA was employed to test for mean differences among the three groups, and the Games-Howell procedure was applied for post hoc analysis. Both analyses were performed using the Pingouin library in Python. First, Welch's ANOVA was used to identify which features showed significant mean differences across groups, and then the Games-Howell post hoc test was conducted to determine the specific group differences for those features. By selecting statistical tests that align with the characteristics of the data at each stage, this study aimed to enhance the validity of the analysis.

4. Results and Discussions

Using the language analysis tool LFTK, the corpora of ESL and EFL learners were analyzed. This study examines written essays of ESL (English as a Second Language) learners such as Hong Kong, Pakistan, the Philippines, and Singapore, as well as EFL (English as a Foreign Language) learners, including China, Indonesia, Japan, Korea, Taiwan, and Thailand. Table 3 summarizes the overall results of both learner groups and English control.

4.1 Vocabulary Profile and Readability

This section aims to provide a general overview of vocabulary use and readability. Table 3 summarizes the average values related to vocabulary and readability, including the total number of words(t_word), total number of unique words(t_uword), total number of sentences(t_sent), average number of words per sentence(a_word_ps),

as well as FKGL[1] and FOGI[2] scores, which are indicators of readability.

Table 3. Average number of words and Readability

	NS	ESL	EFL	*p*-value
t_word	226.80	244.46	232.29	0.000
t_uword	120.33	119.69	111.24	0.000
t_sent	9.41	13.55	15.12	0.000
a_word_ps	25.70	19.52	16.62	0.000
FKGL	10.61	8.60	6.68	0.000
FOGI	14.54	12.41	10.45	0.000

To begin, the results pertaining to vocabulary, which include the 'total number of words,' 'total number of unique words,' 'total number of sentences,' and 'average number of words per sentence' will be examined. An analysis of the average total word count revealed that the ESL had the highest mean at 244.46, followed by the control group (NS) with 226.80, and the EFL with the lowest mean of 232.29. The differences observed between these three groups were statistically significant, showing $p<0.000$, $p<0.000$, $p<0.000$, $p<0.000$ values, respectively. In contrast, when examining the average total number of unique words, the NS and ESL were recorded at 120.33 and 119.69, respectively, with no statistically significant difference between them. However, the EFL exhibited the lowest mean of 111.24, which was found to be statistically significantly lower than both the NS and ESL ($p<0.000$,). The lack of a significant difference between the NS and ESL in terms of unique words suggests that ESL learners are able to utilize a fairly diverse vocabulary, approaching the level of native speakers. The EFL's statistically significant lower vocabulary suggests that their vocabulary range is more limited than ESL, possibly due to less exposure to diverse linguistic contexts and fewer opportunities for vocabulary acquisition.

Next, the average sentence count allows for the calculation of the average number of words per sentence, which can serve as an indicator of sentence length and, consequently, the ability to construct longer, more complex sentences. The average sentence counts across the three groups were 9.41 for the NS, 13.55 for the ESL, and 15.12 for the EFL and average word counts across the three groups were 25.70, 19.52, and 16.62. These differences were found to be statistically significant with $p<0.000$ between all groups, providing valuable insights into the variations in sentence structure and linguistic complexity among the groups. The NS, with an average sentence count of 9.41 and an average word per sentence count of 25.70, tends to produce fewer sentences, containing more words in one sentence.

In contrast, the ESL, with an average sentence count of 13.55 and word per sentence of 19.52, produces significantly more sentences than the NS ($p<0.000$). This higher sentence count may indicate that ESL learners break down their ideas into shorter, more straightforward sentences than NS. These sentences are likely to exhibit simpler syntactic structures compared to those produced by native speakers.

The EFL, with the highest average sentence count of 15.12 and word count 16.62, produces even more sentences than the ESL ($p<0.000$). This increased sentence count further suggests that the EFL group relies on even simpler structures and more frequent segmentation of ideas compared to the ESL. The higher sentence count in the EFL group may reflect a challenge in synthesizing ideas cohesively or constructing more sophisticated sentence structures.

1 FKGL: Flesch-Kincaid Grade Level
2 FOGI: Flesch–Kincaid Readability Index

The differences in total sentence count across the three groups of NS, ESL, and EFL highlight important distinctions in writing style and learning context. While the NS group produces fewer, more complex sentences, both the ESL and EFL groups tend to use a greater number of simpler sentences. These variations show the influence of learning context on writing practices. The higher sentence count in the ESL group compared to the NS group suggests that ESL learners often rely on simpler sentence structures, while the EFL group's even higher sentence count may indicate a greater reliance on basic sentence construction.

Having examined the vocabulary-related aspects, the results concerning readability and the associated indicators will be discussed. The patterns observed in the vocabulary analysis align with the results related to readability. The NS group demonstrates the highest readability, followed by the ESL and EFL groups, with decreasing scores in that order. To further investigate these trends, the FKGL (Flesch-Kincaid Grade Level) and FOGI(Gunning Fog Index) results, a critical indicator of readability will be discussed now.

Firstly, the analysis of FKGL scores among the three groups reveals statistically significant differences in text readability *(*$p < 0.000$ for all comparisons). The English control group exhibited the highest FKGL score (10.61), indicating greater text complexity and a higher level of reading proficiency required. The ESL group had an FKGL score of 8.60, suggesting that their texts, while somewhat less complex, still maintain a moderate level of difficulty. In contrast, the EFL group demonstrated the lowest FKGL score (6.68), indicating that their texts are the most accessible, characterized by simpler vocabulary and shorter sentence structures.

In addition to the FKGL analysis, the FOG index (FOGI) also exhibits a similar trend. The English control group demonstrated the highest FOGI score (14.54), followed by the ESL group (12.41) and the EFL group (10.45) ($p < 0.000$ for all comparisons). These results indicate that the texts produced by the English control group are the most complex in terms of syntactic and lexical demands, whereas those written by the EFL group are the most accessible.

Thus far, the analysis has focused on quantitative metrics related to vocabulary usage and text readability. While these measures provide valuable insights, a more detailed examination of lexical composition is necessary. One key aspect to consider is the total number of unique words, which serves as an indicator of lexical richness and reflects the extent of vocabulary diversity in writing. A higher number of unique words generally suggests greater lexical variation, whereas a lower number may indicate reliance on a restricted set of familiar words. Given the observed differences in vocabulary usage and readability across the three groups, analyzing both the number and composition of unique words will offer deeper insights into how linguistic complexity manifests among learners.

To achieve this, vocabulary will be categorized into unique content words (e.g., nouns, verbs, adjectives, and adverbs) and unique function words (e.g., determiners, auxiliaries, conjunctions, and prepositions). This distinction is crucial for understanding the relationship between lexical usage and readability. Content words convey the core meaning of a sentence and include nouns (e.g., dog, house), verbs (e.g., run, eat), adjectives (e.g., happy, beautiful), and adverbs (e.g., quickly, often). These words are essential for expressing ideas and developing vocabulary, as they carry significant lexical meaning. A learner's ability to use a diverse range of content words reflects their capacity to elaborate on thoughts in written discourse. Function words, on the other hand, primarily serve grammatical roles, such as determiners (e.g., the, a), auxiliary verbs (e.g., is, have), conjunctions (e.g., and, but), and prepositions (e.g., in, on). While they contribute little to semantic content, they are crucial for maintaining syntactic structure and coherence. Their primary function is to connect and organize content within a sentence, ensuring grammatical accuracy and textual cohesion.

Therefore, the NS group's higher readability scores may be associated with a more balanced distribution of content and function words, contributing to syntactic complexity and coherence. In contrast, the lower readability scores of the ESL and EFL groups may reflect differences in how these learners utilize unique vocabulary items. For instance, if the EFL group exhibits a lower proportion of unique function words relative to content words, this

could suggest limited syntactic variety, potentially reducing readability.

By examining both lexical diversity and word composition, this study aims to provide a more nuanced understanding of how ESL and EFL learners structure their writing and how these patterns impact readability. These insights have direct implications for English language education, as they can inform pedagogical approaches tailored to the distinct challenges faced by ESL and EFL learners. A better understanding of their lexical usage can help educators develop targeted strategies to enhance vocabulary development, improve syntactic variety, and ultimately foster more coherent and sophisticated writing

4.2 Analysis on the Content Words and Functional Words

4.2.1 Analysis of student writing

Firstly, let us examine the content words. Linguistic entities that can be regarded as content words category include 'adjective(adj)', 'adverb(adv)', 'noun', and 'verb'. Figure 1 shows the average number of content words by EFL and ESL learners and native speakers.

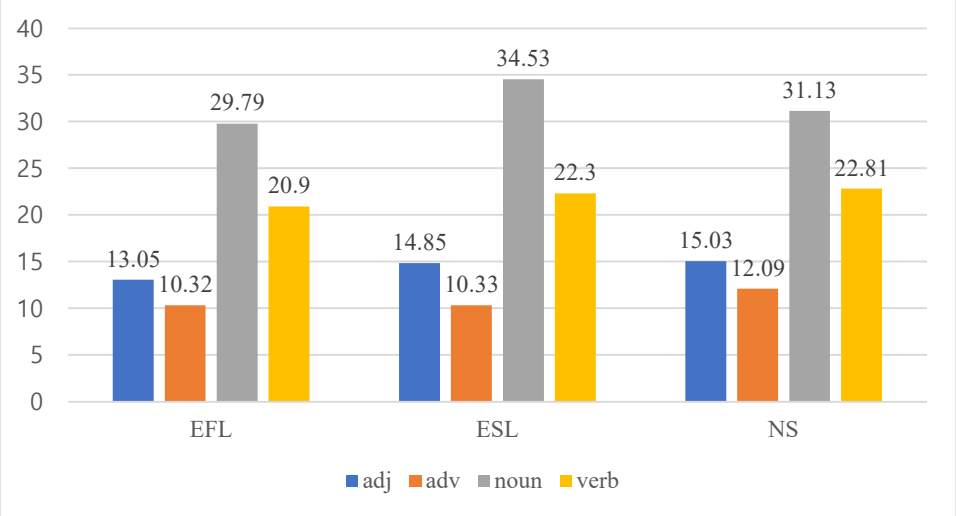

Figure 1. Number of Unique Contents Words

The numbers in Figure 1 represent the number of unique content words including adjective, adverb, noun, and verb. Firstly, in terms of the number of unique adjectives, native speakers exhibited the highest mean (15.03), followed by ESL learners (14.85) and EFL learners (13.05), respectively. Statistical analysis shows a notable difference in the use of unique adjectives between EFL learners and both ESL learners and native speakers (p<0.000, p<0.000, respectively), whereas the difference between ESL learners and NS is not statistically significant.

Regarding the use of adverbs, NS exhibited the highest mean (12.09), while EFL and ESL learners demonstrated lower means of 10.33 and 10.32, respectively. Statistical analysis reveals no significant difference in the use of unique adverbs between EFL and ESL learners. However, both groups demonstrate significant differences when compared to native speakers (p<0.000, p<0.000, respectively). This suggests that EFL and ESL learners employ a more restricted range of adverbs than NS.

In terms of uses of unique nouns, ESL showed the highest mean (34.53) and it is followed by NS (31.13) and

EFL (29.79). All groups exhibit significant differences in their use of unique nouns, with noticeable variations in noun diversity across each group (p<0.000, p<0.000, p<0.000, respectively). These distinct differences highlight a clear gap between native speakers and both EFL and ESL learners in terms of noun variety. The observation that ESL speakers employ a more extensive range of unique nouns compared to native speakers is noteworthy.

For the uses of unique verbs, NS demonstrated the highest mean (22.81), followed by ESL (22.3) and EFL (20.9). The statistical comparison of verb usage across different learner groups reveals significant differences between EFL and ESL learners, as well as between EFL learners and native speakers (NS). However, no significant difference was found between ESL learners and NS. Interestingly, ESL learners appear to use a broader variety of verbs that closely mirrors the patterns observed in native speakers, suggesting that the immersive environment in which they learn may contribute to their more nuanced use of vocabulary. This result show the potential benefits of immersion for mastering the subtleties of verb usage. In contrast, EFL learners, who may not have the same level of exposure to authentic language use, exhibit a more limited range of verbs. Therefore, it is suggested that EFL instruction could be enhanced by placing greater emphasis on expanding learners' verb vocabulary, promoting the use of more varied and contextually appropriate verbs to improve language complexity and fluency.

The comparative analysis of content word usage across native speakers (NS), English as a Second Language (ESL) learners, and English as a Foreign Language (EFL) learners reveals significant discrepancies in vocabulary usage patterns. The results present key differences in the frequency and diversity of content word usage, including adjectives, adverbs, nouns, and verbs, among the three groups.

In general, native speakers demonstrated the highest mean frequency of unique content word usage across all categories, as expected. ESL learners, in turn, exhibited vocabulary usage patterns that closely mirrored those of NS, except for noun usage, where ESL learners used nouns at a higher frequency than NS. Notably, for adjectives and verbs, the mean usage of ESL learners did not differ statistically from that of NS, suggesting that ESL learners are relatively proficient in these areas. In contrast, EFL learners exhibited significant statistical differences in their content word usage compared to NS, indicating that EFL learners exhibit lower lexical diversity compared to native speakers.

4.2.2 Analysis of functional words

This section will report the results of the function word analysis. Functional words include auxiliary verbs, determiners, pronouns, prepositions, coordinating conjunctions, and subordinating conjunctions. The following figure summarizes the mean number of unique vocabulary items in each functional category.

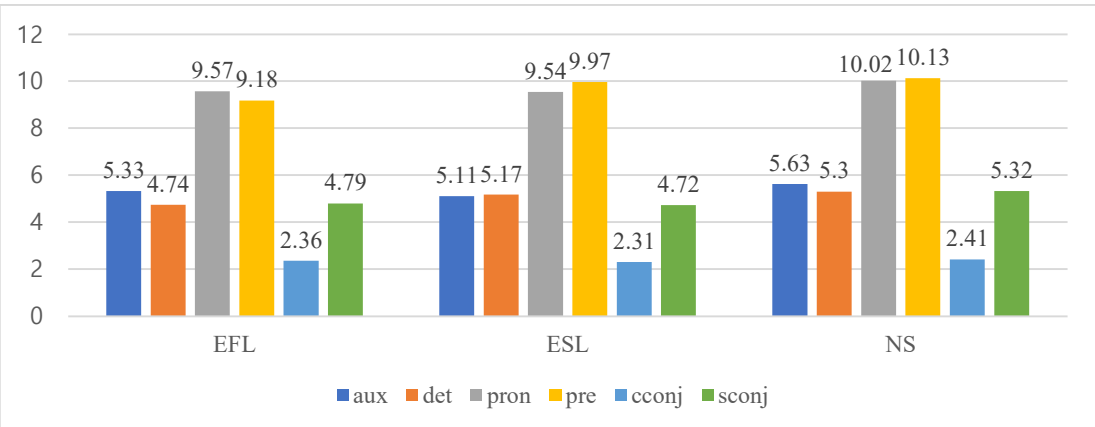

Figure 2. Number of Unique Function Words

Firstly, let us examine the unique auxiliary verb uses. Native speakers demonstrated the highest mean (5.63), with EFL speakers showing 5.33 and ESL speakers showing 5.11. Statistical Analysis shows that significant differences in auxiliary verb usage are evident among all groups, highlighting the varying levels of proficiency. It can be interpreted that achieving diversity in auxiliary verb use poses challenges, even for ESL learners, who may struggle to master this aspect of language despite their exposure to English in diverse contexts. It is also noticeable that while ESL speakers generally demonstrate patterns more similar to native speakers in Content Words, EFL speakers might exhibit more nativelike behavior in auxiliary verb usage.

Regarding the results of unique determiners, NS displayed the highest mean (5.30), followed closely by ESL (5.11), and then EFL (4.74). Statistical analysis reveals no significant difference between NS and ESL learners, suggesting that ESL learners exhibit a similar pattern to native speakers in their use of determiners. In contrast, a statistically significant difference was found between EFL and ESL learners. Although the difference in unique determiner use among NS, ESL, and EFL learners was found to be statistically significant, it is important to consider that the category of determiners is inherently limited in number. Therefore, examining the frequency of determiner usage across the three groups may provide additional insight into learners' grammatical patterns.

The total number of determiners is 18.40 for NS, 21.36 for ESL, and 19.20 for EFL. An interesting phenomenon is that the learner groups exhibited higher frequencies than native speakers, with ESL learners showing the highest frequency, followed by EFL learners and then NS. The differences among all three groups were statistically significant. The result that ESL learners exhibited the highest total number of determiners, followed by EFL learners and native speakers, invites several possible interpretations. One explanation is that ESL learners, who are typically immersed in English speaking environments, may have greater exposure to both input and output opportunities, leading to more frequent use of grammatical elements such as determiners. Interestingly, both ESL and EFL learners produced more determiners than native speakers. This pattern may reflect a tendency among language learners to rely more heavily on explicit grammatical instruction, resulting in a more form-focused style of writing. However, it is important to note that a higher frequency of use does not necessarily indicate grammatical mastery. In both ESL and EFL groups, frequent determiner use may reflect heightened awareness of this grammatical category rather than fully accurate or contextually appropriate application.

In fact, previous studies on article acquisition have shown that once learners become aware of the importance of articles, they often begin to overuse or misuse them, especially in contexts where native speakers would omit them (Ionin et al. 2004, Master 1987, Park 2014). This suggests that increased frequency may be a sign of learners' growing attention to form, but not yet of proficiency. Therefore, frequency alone should not be interpreted as a marker of grammatical competence; rather, it may reflect a transitional stage in interlanguage development for both ESL and EFL learners. The statistically significant differences among all three groups indicate that while learners may approximate native like patterns in certain respects, the quality and appropriateness of their usage remain distinct, underscoring the complexity of mastering this grammatical category.

In examining the use of unique pronouns among different groups, native speakers exhibited the highest mean of 10.02, followed by both EFL learners (9.57) and ESL learners (9.54). The results show no significant differences between EFL learners and ESL learners. However, significant differences are observed between native speakers (NS) and both EFL and ESL learners *(*p<0.000, p<0.000, respectively). This suggests that while EFL and ESL learners share similar pronoun usage patterns, both groups differ notably from native speakers.

Regarding the use of prepositions, NS displayed the highest mean (10.13), closely followed by ESL (9.97) and EFL (9.18). Significant differences were observed in the use of unique prepositions between EFL learners and both ESL learners and native speakers. Specifically, EFL learners showed notable differences in preposition usages when compared to both ESL learners and native speakers. However, no significant differences were found between

ESL learners and native speakers. However, no significant differences were found between ESL learners and native speakers, which suggests that ESL learners, possibly as a result of their greater exposure to English in immersive environments, tend to use prepositions in a manner comparable to that of native speakers. The classification of prepositions has been the subject of ongoing debate across various linguistic frameworks. Traditionally, prepositions have been categorized as function words due to their grammatical role in linking noun phrases to other sentence elements (Huddleston and Pullum 2002, Quirk et al. 1985).

In generative grammar, they are treated as functional categories that occupy specific syntactic positions, with emphasis placed on structural relationships rather than semantic properties. However, from a cognitive linguistic perspective, prepositions are viewed as semantically rich elements that reflect how speakers conceptualize space, time, and abstract relations. Scholars such as Tyler and Evans (2003) have argued that prepositions encode image schemas and undergo semantic extensions, suggesting that their meanings are best understood in terms of prototypical senses and conceptual networks. This view challenges the traditional notion of prepositions as merely grammatical markers, proposing instead that they occupy a unique position between function and content words. In second language acquisition research, prepositions are widely recognized as one of the most difficult grammatical categories for learners to master (Celce-Murcia and Larsen-Freeman 1999, Tyler and Evans 2003). Errors in preposition use are often attributed to negative transfer from the first language, overgeneralization, and limited exposure to authentic contexts. Moreover, studies have shown that once learners become aware of the grammatical importance of prepositions, such as spatial markers, they tend to overuse or misuse them (e.g., in, on, at) (Master 1997, Tyler and Evans 2003). These results reveal the dual nature of prepositions as both structurally functional and semantically loaded, and they highlight the need for instructional approaches that address not only form but also conceptual meaning.

In terms of uses of unique coordinating conjunctions, NS showed the highest mean of 2.41 which are closely followed by EFL (2.36) and ESL (2.31). In analyzing the use of unique coordinating conjunctions, no significant differences were observed among the groups. This suggests that EFL learners, ESL learners, and native speakers (NS) all exhibit similar patterns in their use of coordinating conjunctions. Despite differences in other areas of language use, the use of coordinating conjunctions appears to be relatively consistent across these groups, indicating that this aspect of language may be less influenced by language learning context or exposure. Alternatively, it is also possible that this phenomenon occurs because the number of coordinating conjunctions available for use under the given topic can be inherently limited. Assuming that the number of coordinating conjunctions is inherently limited, we analyzed the average usage frequency of this grammatical feature among native speakers (NS) and learners. The results indicate that NS demonstrates the highest frequency of coordinating conjunction usage with mean number of 8.63, followed by ESL learners (8.15), with EFL learners (7.28) showing the lowest frequency. These differences in usage frequencies are statistically significant, highlighting the variation in coordinating conjunction use across the groups.

Regarding the use of unique subordinating conjunctions, NS showed the highest mean (5.32), and followed by EFL (4.79) and ESL (4.72). The use of unique subordinating conjunctions reveals distinct patterns across different groups. No significant difference was found between EFL and ESL learners, suggesting similar levels of subordinating conjunction usage in these two groups. However, both EFL and ESL learners differ significantly from native speakers (NS), with the latter demonstrating a broader variety of subordinating conjunctions. This lower variety of both conjunctions among EFL and ESL learners may reflect a tendency to rely on simpler sentence structures, potentially due to limited exposure to more complex syntactic forms. Emphasizing the use of subordinating conjunctions in language instruction could help learners develop more sophisticated sentence structures, thereby enhancing their overall proficiency in constructing complex and cohesive sentences.

The present study has examined the grammatical features reflected in the number of unique lexical items within both content and function word categories. Based on the analysis, the results may be summarized as follows. Overall, both EFL and ESL learners exhibited a significant difference in the use of function words compared to native speakers. However, it is notable that while the average use of content words in most categories was closer to that of native speakers for ESL learners than for EFL learners, in several function word categories such as auxiliary verbs, pronouns, coordinating conjunctions, and subordinating conjunctions, the usage patterns of EFL learners more closely resembled those of native speakers. In the following section, pedagogical strategies appropriate for ESL and EFL learners will be proposed based on these results.

5. Pedagogical Implications

To briefly summarize, ESL learners appear to demonstrate patterns in their use of content words that are more closely aligned with those of native speakers, as compared to their EFL counterparts. This tendency seems to reflect a broader and more diverse use of lexical items within the categories of nouns, verbs, adjectives, and adverbs. Such results may show that learners situated in immersive language learning environments are more likely to benefit from frequent and varied exposure to content rich vocabulary. In light of these observations, it may be suggested that EFL learners could similarly benefit from instructional approaches that emphasize meaningful and varied vocabulary input, especially those that incorporate elements of immersion whenever feasible.

Building upon this suggestion, the integration of AI-based vocabulary learning tools, such as intelligent tutoring systems (e.g., Rosetta Stone, Babbel), adaptive vocabulary platforms (e.g., Quizlet, Memrise), and conversational agents (e.g., Duolingo, Replika), emerges as a promising pedagogical tools for promoting lexical diversity in EFL contexts. In settings where access to immersive environments is limited, these AI technologies can simulate elements of immersion by offering learners personalized and meaningful exposure to content words. Specifically, intelligent tutoring systems like Rosetta Stone provide immediate, individualized feedback on learners' vocabulary usage, correcting errors in real time and suggesting more contextually appropriate alternatives. Adaptive vocabulary platforms such as Quizlet and Memrise tailor the learning experience based on the learner's proficiency level, presenting vocabulary that is both challenging and relevant to their current linguistic abilities. Conversational agents, such as Duolingo and Replika, engage learners in dynamic interactions, encouraging the use of newly learned vocabulary in realistic communicative contexts.

Through repeated exposure to a wide range of lexical items, these tools can help learners reinforce their understanding of word meanings and their correct usage across different contexts. Furthermore, AI platforms often incorporate spaced repetition algorithms to optimize retention and long-term learning. Recent studies have demonstrated that AI mediated instruction not only supports vocabulary acquisition but also enhances learner motivation and self-regulated learning (Alsakaker 2025, Azizollahi et al. 2025, Ma and Chen 2024, Wei 2023). Incorporating such technologies into classroom practice may thus offer an effective means of addressing the lexical development gap observed between EFL and ESL learners.

An analysis of the functional category words reveals that the results can be interpreted as an indicator of how complex sentence structures each learner group is able to produce. Unlike content words, where there was little variation across the groups, the use of functional words showed more distinct patterns. In particular, when compared to the functional category, EFL learners exhibited language patterns that were closer to those of native speakers than ESL learners. This was particularly evident in the use of auxiliary verbs, pronouns, coordinating conjunctions, and subordinating conjunctions. Such patterns may reflect the outcomes of form-focused instruction

commonly found in EFL learning environments, where greater attention is given to the structural aspects of language. This suggests that EFL learners, through targeted instruction, may develop a higher degree of proficiency in constructing grammatically complex sentences, which mirrors some of the syntactic behaviors seen in native speakers.

Although ESL learners are typically immersed in English speaking environments, the results of this study present that EFL learners demonstrated patterns in the use of functional words that were more similar to those of native speakers. This unexpected result suggests that immersion alone may not be sufficient for developing accurate and complex grammatical structures, highlighting the potential value of form-focused instruction in ESL contexts. ESL learners may therefore benefit from more explicit instruction targeting the use of functional words, particularly auxiliary verbs, pronouns, and both coordinating and subordinating conjunctions, which play a critical role in syntactic complexity and overall fluency.

Incorporating targeted grammar-focused activities within communicative and content-based ESL instruction could help bridge the gap between natural exposure and grammatical accuracy. For instance, guided writing tasks, structured speaking activities, and focused grammar workshops can provide ESL learners with opportunities to reflect on and refine their use of functional language elements. Ultimately, these results point to the need for a more balanced instructional approach in ESL environments that combines the strengths of immersion with the benefits of explicit grammar instruction. By integrating form-focused components into an otherwise meaning-oriented curriculum, educators can help learners achieve greater syntactic control and closer alignment with native-like language use.

This implication is further supported by Schenck's (2019) meta analysis, which demonstrated the effectiveness of explicit grammar instruction in improving learners' production accuracy. The study emphasized that explicit focus on grammatical features, particularly those dissimilar from learners' L1, can lead to significant gains in accuracy. It can be suggested that even in immersive ESL contexts, integrating explicit instruction can enhance learners' control over functional language use. Therefore, in ESL education, it is essential to appropriately balance explicit and implicit instructional methods by considering factors such as L1 similarity, grammatical complexity, and learners' proficiency levels. A tailored approach of this kind can enhance both grammatical accuracy and overall language proficiency.

6. Conclusion

In conclusion, this comparative analysis highlights the differences in lexical development between EFL learners and ESL learners, with important pedagogical implications. The results imply that while ESL learners benefit from immersive environments that foster greater lexical flexibility and content word diversity, EFL learners tend to exhibit patterns in the use of functional words that are more similar to those of native speakers, likely as a result of form-focused instruction. Integrating AI based vocabulary learning tools can help EFL learners simulate immersive exposure, promoting lexical diversity and motivation even in non-immersive contexts. Conversely, ESL learners may benefit from explicit instruction targeting functional words to enhance grammatical complexity and accuracy. Ultimately, a balanced instructional approach that combines meaningful vocabulary input with targeted grammar instruction, tailored to learners' contexts and needs, is essential for bridging the lexical and syntactic gaps between EFL and ESL learners and supporting their progression toward greater similarity with native speaker language use.

Although previous research has rarely compared EFL and ESL learners using a large-scale corpus such as

ICNALE, this study fills that gap by showing how learning contexts shape differences in content and function word use. This unique contribution advances theoretical understanding of lexical development across contexts and offers practical implications for pedagogy.

Despite these contributions, several limitations should be acknowledged. First, the analysis focused primarily on written texts, which may not fully represent learners' vocabulary use across other modalities such as speaking or listening. Second, although the study compared groups based on their learning contexts, individual differences such as motivation, exposure to English outside the classroom, and educational background were not controlled for, which may have influenced the results. Third, while some observations were made regarding functional word use, the study did not conduct a comprehensive syntactic analysis. Expanding future research to examine properties of broader syntactic categories could offer deeper insights into persistent grammatical difficulties and inform more targeted instructional strategies. Such work would build upon the present study's pedagogical implications, further supporting the development of balanced and context-sensitive approaches to language instruction.

References

Alsakaker, S. M. 2025. Investigating EFL Learners' perceptions of using AI to enhance English vocabulary acquisition based on the technology acceptance Model. *Forum for Linguistic Studies* 7(2), 1067-1077.

Azizollahi, Z., F. Tabrizi, and Z. Behineh. 2025. Utilizing AI mediated methods of EFL vocabulary learning and teaching. *Modarese Bartar Quarterly* 2(3), 1-18. Available online at https://teacherstribune.com/wp-content/uploads/2025/01/Utilizing-AI-Mediated-Methods-of-EFL-Vocabulary-Learning-and-Teaching-1.pdfArticle Publication

Azodi, N., F. Karimi and R. Vaezi. 2014. Measuring the lexical richness of productive vocabulary in Iranian EFL university students' writing performance. *Theory and Practice in Language Studies (TPLS)* 4(9).

Bestgen, Y. 2017. Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System* 69, 65-78.

Bulté, B. and A. Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26, 42-65.

Buschfeld, S. and A. Kautzsch. 2017. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* 36, 104-126.

Celce-Murcia, M. and D. Larsen-Freeman. 1999. *The Grammar Book: An ESL/EFL Teacher's Course* (2nd ed.). Heinle and Heinle.

Chang, X. and P. Wang. 2016. Influence of second language proficiency and syntactic structure similarities on the sensitivity and processing of English passive sentence in late Chinese-English bilinguists: an ERP study. *Journal of Psycholinguistics Research* 45, 85-101.

Ma, Y. and M. Chen. 2024. AI-empowered applications effects on EFL learners' engagement in the classroom and academic procrastination. *BMC Psychology* 12, 739

Djiwandono, P.I. 2016. Lexical richness in academic papers: A comparison between students' and lecturers' essays. *Indonesian Journal of Applied Linguistics* 5(2), 209-216.

Engber, C.A. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4(2), 139-155.

González-López, S. and A. López-López. 2015. Lexical analysis of student research drafts in computing. *Computer Applications in Engineering Education* 23(4), 638-644.

Ha, D. 2019. Reinforcement learning for improving agent design. *Artificial Life* 25(4), 352-365.

Huddleston, R. and G. Pullman 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Ionin, T., H. Ko and K. Wexler. 2004. Article semantics in L2 acquisition: The role of specificity. *Language acquisition* 12(1), 3-69.

Ishikawa, S. 2011. ICNALE: The international corpus network of Asian learners of English. *Corpora* 6-2, 121-133.

Ishikawa, S. 2011. A corpus-based study on Asian learners' use of English linking adverbials. *Themes in Science and Technology Education 3* 1-2, 139-157.

Ishikawa, S. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa, ed., *Learner corpus studies in Asia and the world* 1, 91-118. Kobe University.

Ishikawa, S. 2024. Lexicogrammar of the L2 English essays written by Asian college students: a corpus-based Study. *Journal of Asia TEFL* 21(1), pp.119-138.

Ishikawa, S. 2025. Quantitative reconsideration of an L1 effect on Asian learners' L2 English writing: a study based on the ICNALE. *LEARN Journal: Language Education and Acquisition Research Network* 18(1), 989-1014.

Johnson, M. D., L. Mercado, and A. Acevedo. 2012. The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing* 21, 264–282.

Krashen, S. D. 1985. *The Input Hypothesis: Issues and Implication*. Longman.

Kyle, K. and S.A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4), 757-786.

Laufer, B. and P. Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics* 16(3), 307-322.

Lee, B. W. and J. Lee. 2023. *LFTK: Handcrafted features in computational linguistics*. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 1-19). Association for Computational Linguistics.

Loewen, S. 2015. *Introduction to Instructed Second Language Acquisition.* Routledge.

Long, M.H. 1981. INPUT, INTERACTION, AND SECOND-LANGUAGE ACQUISITION. *Annals of the New York Academy of Sciences* 379, 259-278.

Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15, 474–496.

Lu, X. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal* 96(2), 190-208.

Master, P. 1987. Teaching the English article as a binary system. *TESOL Quarterly* 24(3), 461-478.

Nation, I.S. and I.S.P. Nation. 2001. *Learning Vocabulary in Another Language*. Cambridge university press.

Norris, J. M. and L. Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics* 30, 555-578.

O'Leary, J.A. and R. Steinkrauss. 2022. Syntactic and lexical complexity in L2 English academic writing: development and competition. *Ampersand 9,* 100096.

Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics* 24(4). 492-518.

Ortega, L. 2005. Methodology, epistemology, and ethics in instructed SLA research: An introduction. *The Modern Language Journal* 89(3). 317-327.

Ortega, L. 2015. Syntactic complexity in L2 writing: progress and expansion. *Journal of Second Language Writing* 29, 82-94.

Park, S. 2014. *L2 Acquisition of Genericity in English Articles: The Case of Korean Adult Learners of L2 English.* Doctoral dissertation. University of Sheffield.

Park, S. 2023 Corpus analysis of L2 English article usage patterns and pedagogical implications. *Cogent Education* 10(1).

Quirk, R., Greenbaum, S., Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English language*. Longman.

Schenck, A. 2019. The impact of form-focused instruction on Korean learner production: A comprehensive study of technique and timing. *English Teaching* 74(2), 75-102.

Schmitt, N. 2008. Instructed second language vocabulary learning. *Language Teaching Research* 12(3), 329-363.

Shehadeh, A. and A. Coombe. 2012. *Task-Based Language Teaching in Foreign Language Contexts: Research and Implementation*. Task-Based Language Teaching: V. 4. John Benjamins Pub. Co.

Tyler, A. and V. Evans. 2003. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meanings and Cognition*. Cambridge University Press.

Vedder, I. and V. Benigno. 2016. Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching* 54(1), 23-42.

Vold, E. T. 2022. Learner spoken output and teacher response in second versus foreign language classrooms. *Language Teaching Research* 29(2).

Wei. L. 2023. Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology.*

Zhang, Y. and J. Ouyang. 2023. Linguistic complexity as the predictor of EFL independent and integrated writing quality. *Assessing Writing* 56, 100727.

Examples in: English
Applicable Languages: English
Applicable Level: Primary/Secondary/Tertiary

