International Journal of English Language and Pedagogy (IJELP)
Vol. 3, No. 1, pp. 109 - 126
EISSN: 2964-6545, DOI: 10.33830/ijelp.v3i1.12042

Variation in Linguistic Complexity and Human Ratings of Second Language (L2) Writing

Thanh T.G. Trinh
Faculty of Foreign Languages, Ho Chi Minh City University of Technology and Education

Article Info

Article history:

Received: February 20th, 2025
Revised: April 3rd, 2025
Accepted: May 14th, 2025

Keywords:

Inter-individual variation
linguistic complexity
usage-based theory
Second language writing

ABSTRACT

Drawing upon the usage-based approach, which considers inter-individual variation as an inherent feature of language use and production, this study attempted to explore what linguistic complexity measures significantly differed among individual learners of similar learning conditions and how they were compatible with rubric-based scorings. To achieve this, a corpus of 56 academic essays written by five upper-intermediate Vietnamese learners enrolled in an IELTS preparation course was analyzed. Linguistic complexity was operationalized across lexical, syntactic, and discoursal domains using a wide range of validated computational tools, including TAALES, TAALED, TAASSC, and TAACO. Statistical analyses employed non-parametric tests to identify significant inter-learner differences and correlations with holistic human ratings based on official IELTS rubrics. Inter-rater reliability was also assessed to ensure scoring consistency. The findings confirm that variation in complexity ranges from lexical, syntactic, to discoursal levels. Comparisons with human ratings reveal that the variation in linguistic complexity measures is complex in nature since complexity correlates with perceived proficiency, particularly in semantic cohesion and lexical control, but in other cases varies independently of performance, suggesting that the developmental stage of complexity may moderate its impact on ratings.

Corresponding Author:

Thanh T.G. Trinh
Email: thanhttg@hcmute.edu.vn

1. INTRODUCTION

A usage-based approach conceptualizes language development as emerging from meaningful, repeated use in communicative contexts (Tomasello, 2003; Langacker, 2008; Verspoor & Behrens, 2011). Accordingly, L2 learners acquire and make linguistic choices through exposure to input, practice, and social interaction, leading to the gradual entrenchment of form-function mappings that support effective discourse production. In other words, from this perspective, language production is individually owned and contextually driven, leading to an undeniable fact that inter-

individual variation is an inherent property of L2 production. Moreover, in recent years, variation in linguistic complexity has become of dominance in the SLA research strand which places emphasis on developmental variation (Kuiken et al., 2019). Ortega (2012) pointed out that the investigation in linguistic complexity should move beyond isolated syntactic measures to encompass a more comprehensive and dynamic understanding of how different dimensions of complexity interact and develop in second language acquisition.

With the rapid development of natural language processing (NLP) tools, the computational measures for linguistic complexity have exponentially grown, attempting to assess the complexity of a text in a short time. Grounded in usage-based theory, this study attempts to explore and explain the variation in linguistic complexity of L2 academic writing, employing these computational measures. In this study, the emphasis lies in the assumption that language use and production are inherently individualized processes, shaped by each learner's unique trajectory of internalization. Even under similar instructional conditions, learners process, practice, and produce language differently, resulting in distinct patterns of linguistic complexity.

The following section presents the theoretical background on which this study is developed and summarize the computational linguistic complexity measures which have been substantially employed in L2 writing research.

2. LITERATURE REVIEW

Literature review is a comprehensive investigation of the available theoretical background including from books and scholarly articles related to your research areas and theories. In this section, you should provide a description, summary, and critical evaluation of your works concerning the research problem being investigated. Literature reviews are aimed at providing an overview of sources you have explored while researching a particular topic to notify your readers how your research fits within a larger field of study.

1.1. Linguistic complexity in L2 writing

There is no agreed-upon definition of complexity; however, numerous SLA research has proven that complexity is "multilayered, multifaceted and multidimensional in nature" (Kuiken, 2023, p. 84). According to Bulté and Housen (2014), linguistic complexity is generally interpreted as a quantitative property with respect to the number and range of linguistic elements in an entity or system. Linguistic complexity can be regarded as an important indicator of L2 performance, reflecting language proficiency and language development (Verspoor et al., 2017). In L2 writing research, linguistic complexity is typically analyzed through complexity measures of the lexicon, syntax, phraseology, and text discourse (Kyle, 2016; Graesser et al., 2004; Wolfe-Quintero et al., 1998).

2.1.1. Lexical complexity

Lexical complexity is a multifaceted construct comprising three interrelated subcomponents: lexical sophistication, lexical density, and lexical diversity. (Kim et al., 2018; Michel, 2017; Lu, 2012).

2.1.1.1. Lexical sophistication

Lexical sophistication, or rareness, refers to the percentage of sophisticated or difficult words in a text (Laufer & Nation, 1995). Traditionally, this domain was measured by VocabProfilers to reveal the frequency in which a word occurs (Cobb, 2009). According to Schmitt (2010), the frequency of sophisticated or difficult words "is arguably the single most important characteristic of

lexis that researchers must address" (p. 63). Meanwhile, substantial research employs average word length (AWL) as an index of lexical sophistication (Grant & Ginther, 2000;Verspoor et al., 2012, Verspoor et al., 2017) since the length of academic or less frequent words are generally greater than that of frequently used words. AWL is believed to reflect lexical sophistication by means of both length and frequency. However, recent computational linguists posit that lexical sophistication encompasses various dimensions beyond frequency. Through Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015), other lexical sophistication domains such as orthographic density, association strength, and so on were included (Kim et al., 2018).

2.1.1.2. Lexical density

Lexical density is the ratio of the number of lexical or content words over the total number of words in a text (Ure, 1971; Fang & Pace, 2013). The higher the lexical density index is, the more information-carrying words text contains, entailing greater complexity (Lankshear & Knobel, 2004; Johansson, 2009). Lexical density can be obtained by employing the Automatic Analysis of Lexical Diversity (TAALED; Kyle et al., 2021) for both type and token indices. Accordingly, lexical density by types is calculated by content word types (N) divided by word types (N) whereas lexical density by tokens is computed by content word tokens divided by word tokens.

2.1.1.3. Lexical diversity

Lexical diversity (variation or richness) is an indicator of vocabulary range and variety in a text (Tweedie & Baayen, 1998) and was traditionally measured by the type-token ratio (TTR) as the ratio of the number of different words (types) to the total number of words (tokens) (Read, 2000). Since TTR is highly affected by text length (McCarthy & Jarvis, 2007), language development studies also resort to Guiraud, a rooted TTR. The advantage of Guiraud is that the curve tends to be flattened, resulting in more observable differences in the diversity of lexical items (van Hout & Vermeer, 2007; Verspoor et al., 2012, Trinh et al., 2023). Recently, more advanced methods, e.g. TAALED (Kyle et al., 2021) have been introduced to incorporate a wider variety of lexical diversity measures.

2.1.2. Syntactic and phrasal complexity

Syntactic complexity quantitively measures the degree of complexity of the syntactic constructions used in L2 production. Common indices of syntactic complexity can be based on length (e.g., average sentence length), number of constituents of syntactic units (e.g., modifiers per noun phrase) or the ratio or density of particular types of structures (e.g., complex T-unit ratio, prepositional phrases per 1000 words) (Graesser et al., 2004; Wolfe-Quintero et al., 1998). In examining syntactic complexity, these measures are commonly classified into large-grained and fine-grained indices depending on the degree of specificity of the types of structures they gauge.

For holistic syntactic complexity measures, syntactic complexity are conventionally measured across its numerous dimensions, including global complexity, complexity by subordination, coordination, and complexity via phrasal elaboration, as suggested by Norris and Ortega (2009). The L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010; Ai & Lu 2013; Lu & Ai 2015), a free python-based automated text analyzer can yield the data of 14 large-grained syntactic complexity indices across five different categories of syntactic.

Regarding fine-grained syntactic complexity and sophistication, the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016) provides a large repertoire of measures at clausal and phrasal levels. The indices of clausal complexity generally gauge the number of specific structures per clause. However, clausal complexity is found to be

neither a strong indicator of essay quality (Kyle, 2016) nor a characteristic feature of academic writing (Biber et al., 2014). On the contrary, previous research highlights the significance of phrasal complexity measures related to complex noun phrases, nominals, embedding information in noun phrases, etc. as critical determiners of writing proficiency (Kyle, 2016; Crossley & McNamara, 2014; Guo et al., 2013; McNamara et al., 2010). In TAASSC, phrasal complexity is reflected multidimensionally by seven types of noun phrase (e.g., nominal subject) and ten types of phrasal dependents (e.g., adjectival modifiers).

Syntactic sophistication refers to how complex the syntactic structures used in L2 production are. TAASSC (Kyle, 2016) provides a set of usage-based indices based on the verb-argument construction (VAC) and main verb lemma frequencies as well as the strength of association between different verb lemmas and VACs (Lu, 2025). From a usage-based perspective, constructions that are more frequent in the input will be learned earlier/more easily (Tomasello, 2003) than less frequent constructions. Therefore, better L2 writing means the ability to use less frequently encountered constructions. Apart from frequency, association strength has drawn attention of usage-based researchers who have found the links between writing quality and association strength of verb-VAC combinations. For example, Kyle (2016) found that essays with weakly associated verb-VAC combinations earned lower scores, while essays with strongly associated verb-VAC combinations achieved higher scores. Another significant finding was that L2 English learners of higher proficiency level produced similar patterns of verbs in VACs to native speakers (Römer et al., 2014; 2015; Römer & Garner, 2019).

2.1.3. Discoursal complexity

Discoursal complexity is composed of multiple components, one of which is textual cohesion. Textual cohesion is defined as the degree to which ideas within a text are connected through the uses of cohesive devices. In this section, the distinction between cohesion and coherence will be briefly described, which is followed by a discussion about the types of cohesive devices commonly used to measure text discourse complexity.

Cohesion and coherence of a text are interconnected elements which a writer employs to convey ideas or communicate within and beyond sentence level. There seems to be a blurring distinction between cohesion and coherence because of their dependence on linguistic elements their meaning-making ability. Coherence is a means for the writer to persuade and interact with the reader; therefore, helps the reader with text interpretation (Malmkjaer, 2001; Reynolds, 2001; Sanders & Maat, 2006). The function of cohesion is to glue different parts of a text logically (Alarcon & Morales, 2011; Salkie, 1995). While the focal point of the former is how ideas are logically structured to form meaningful and comprehensible text, the latter focuses on how lexical and grammatical devices are employed. In other words, coherence implies the text comprehensibility in the mind of the reader whereas cohesion is in relation to the utilization of lexico-grammatical devices in a text (Bui, 2022; Crossley et al., 2016).

Cohesive devices may be utilized at local, global and textual levels (Graesser et al., 2004). Local cohesive devices refer to the overlaps of lexical items and concepts and the uses of explicit connectives between sentences (Halliday & Hasan, 2014). Global cohesive devices include semantic and lexical overlaps between paragraphs in a text, measuring to what extent words or ideas in one paragraph are repeated in subsequent paragraphs (Foltz, 2007). Text cohesion devices include givenness and causal relations which occur throughout an entire text (Graesser et al., 2004; Halliday & Hasan, 2014). In general, local cohesion is more explicitly found than global and text cohesion.

A commonly used tool to analyze cohesion is Coh-Metrix (Graesser et al., 2004). However, the limitations of this tool lie in its restricted number of indices and batch processing. The Tool for the Automatic Analysis of Cohesion (TAACO; Crossley et al., 2016, 2019) provides a wider range of cohesive indices of all three levels and are open to batch processing, including groups of connective-based indices, of between-sentence or between-paragraph lexical overlap, of between-sentence or between-paragraph semantic overlap or similarity, and four givenness indices.

1.2. L2 writing complexity research from usage-based perspective

2.2.1. Usage-based approach to L2 writing

Usage-based theories conceptualize language learning as a product of language use, which argues that linguistic competence is not innately pre-specified but emerges gradually from meaningful exposure to and repeated engagement with language in authentic contexts (Tomasello, 2003; Langacker, 2008; Verspoor & Behrens, 2011). Core constructs within this framework include frequency of input (Larsen-Freeman, 1976), salience of forms, and the entrenchment of constructions through usage. Learners extract patterns, i.e. form-meaning pairings or constructions based on their frequency and communicative relevance, leading to abstraction and generalization over time. In this view, many variables such as L1, age, motivation, type of exposure, intelligence or context play a central role, rather than domain-specific grammatical modules. Thus, language acquisition is both social and experiential, driven by use rather than rule memorization. As Verspoor et al. (2012) highlights, “one would expect differences among L2 learners, resulting in diverse individual trajectories and plenty of trials and errors along the way” (p. 241).

Second language writing, as a cognitively demanding and socially situated activity, involves the retrieval and recombination of learned constructions. As learners engage with written texts through instruction, reading and writing, they are exposed to recurring linguistic patterns that are gradually internalized. Scholars such as Verspoor et al. (2012), Biber et al. (2004), Ellis et al. (2016) emphasize that written proficiency emerges not from isolated rule learning but from the cumulative effect of processing, practicing, and producing language in authentic tasks. Usage-based theory highlights how writing development is grounded in usage frequency, contextualized input, and repeated output.

2.2.2. Empirical L2 writing studies from usage-based perspective

One of the main research strands of L2 writing complexity is drawn upon the perspective of Complex Dynamic Systems Theory. Studies in line with this strand adopt longitudinal designs to trace the complex, iterative developmental trajectories of linguistic complexity based on the assumption that inter- and intra-learner variability is an inherent property of L2 writing development (e.g., de Bot et al., 2007; Larsen-Freeman, 2017, Verspoor et al., 2008; Vyatkina, 2012, 2013). Several studies are, to some extent, aligned with the current research.

Firstly, Verspoor et al. (2012)’s cross-sectional study explored objective measures for evaluating written texts of 437 L2 learners from Dynamic Usage-Based (DUB) perspective. The texts were holistically rated for proficiency, from beginner to intermediate and 64 linguistic variables at the sentence, phrase, and word levels were analyzed. The results showed that broad, frequently used measures (i.e. sentence length, the Guiraud index, overall use of dependent clauses, multi-word expressions, etc.) effectively differentiated proficiency levels, aligning with previous findings. In line with a dynamic usage-based view, most specific, fine-grained constructions exhibited non-linear developmental patterns, considerable variation, and evolving interrelations among variables.

The most striking illustration of variation as a natural feature of language development comes from Chan et al. (2015), who tracked identical twins with the same teacher and equal exposure to English over a year, using identical tasks in both spoken and written modes, yet found that the twins

exhibited distinct developmental patterns even in broad lexical and syntactic complexity measures. Another relevant research is Verspoor et al. (2017)'s longitudinal study found out that FVR and AWL were the best general measures of linguistic complexity. The three learners share similar developmental trajectories in FVR and AWL. However, at clausal complexity, their development was different. Moreover, O'Leary & Steinkrauss (2022) took interest in lexical and syntactic complexity of intermediate L2 learners' academic essays. However, the focus was to examine how the sub-systems of lexical and syntactic complexity correlated. One important study by Zhang & Zhang (2023) examined the development of lexical cohesion and also concluded that Unlike lower proficiency or beginner learners, intermediate learners are more capable of using global lexical cohesion to link ideas in their writing, as their higher proficiency may free up cognitive resources, allowing them to concentrate on creating a cohesive and coherent text through effective lexical connections.

2.3. Summary and Research gaps

Recent advances in corpus linguistics and computational analysis have enabled researchers to operationalize linguistic complexity in writing through quantifiable indices. Importantly, they also allow researchers to trace individual patterns, making them especially relevant for studying learner differences in usage-based development.

Earlier L2 research has focused on syntactic and lexical forms of complexity. This has led to "a rather narrow, reductionist, perhaps even simplistic view on and approach to what constitutes L2 complexity" (Bulté & Housen, 2012, p.34). Kuiken (2023) suggests linguistic complexity be investigated together with other language proficiency constructs, calling for an adoption of a broader perspective on linguistic complexity. Accordingly, the operationalizations of linguistic complexity should be extended beyond syntactic or lexical complexity as earlier L2 research has focused on. Similarly, Lu (2025), in his conceptual review article, points out that discoursal complexity has received less attention than lexical, syntactic, and phraseological complexity. Recently, discoursal complexity has been included as an additional dimension of linguistic complexity. As Crossley et al. (2016) highlighted, future studies may consider how cohesion interacts with other linguistic elements such as the lexicon and syntax in explaining growth and predicting writing quality. To my knowledge, none of the studies on L2 writing collectively and simultaneously look at lexical, syntactic and discoursal levels.

Another obvious research gap is the learner corpora on which linguistic complexity research has been conducted. The majority of previous studies which have been reviewed above touch upon the L2 writing performance and proficiency of English learners in China, the USA or some countries in Europe. No Vietnamese learner corpus of writing samples has been brought into analysis in terms of linguistic complexity.

Addressing these gaps, the current study focuses explicitly on three key linguistic complexity dimensions - grammatical range (syntactic complexity), vocabulary range (lexical complexity), and coherence and cohesion (discourse complexity), employing a wide range of computational indices. Adopting a case study design, this research employs a corpus of 56 writing samples by 05 Vietnamese learners in an IELTS Preparation Course in order to compare the performances in linguistic complexity among these learners and to explore how these complexity indices correlate with human ratings based on the officially published IELTS Task 2 writing rubric.

Research questions:

1. To what extent are academic essays written by upper-intermediate learners similar or different in terms of lexical complexity?

2. To what extent are academic essays written by upper-intermediate learners similar or different in terms of syntactic complexity?
3. To what extent are academic essays written by upper-intermediate learners similar or different in terms of discoursal complexity?
4. To what extent are these linguistic complexity dimensions compatible with the rubric-based holistic ratings?

3. METHOD

3.1. Context and participants

In Vietnam, high-stakes English language proficiency tests, such as the International English Language Testing System (IELTS), play a significant role as gate-keeping mechanisms in academic and professional contexts worldwide (Pearson, 2019). Specifically, IELTS has been employed for a wide range of purposes, such as academic admissions, university graduation as well as job opportunities, influencing Vietnamese undergraduate students.

The data of the current study were collected from 05 Vietnamese students (4 female, 1 male) enrolled in an IELTS Writing preparation course. They did not have any prior experience with IELTS. The entry test for the course revealed no disparity in their overall English skills. At the beginning of the course, they were assessed to have upper-intermediate English proficiency level and also expressed their ambition to achieve an IELTS bandscore of 7.0 or above. The course lasted about 4 months, with two 2-hour class meetings per week.

The learner corpus included 56 IELTS Task 2 essays covering common writing topics of this international test, i.e., traditions and cultures, crime, tourism, studying abroad, technology, etc. The tutor of the enrolled course gave instructions on how to outline and deal with each task type: discursive essays, opinion essays, argumentative essays, etc. The students wrote the essays as their homework and no corrective feedback was made. The total number of words was 18,741.

Table 1. Descriptions of learner corpus

Student	Gender	Number of essays written	Total number of words	Min	Max
Student 1	Female	11	3,026	239	319
Student 2	Female	09	3,072	247	521
Student 3	Female	12	4,341	279	463
Student 4	Female	13	4,455	309	373
Student 5	Male	11	3,847	293	456

3.2. Computational complexity measures: Selection and procedures

3.2.1 Lexical complexity:

In this study, I employed the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) and the Tool for the Automatic Analysis of Lexical Diversity (TAALED; Kyle et al., 2021) to obtain a vast array of lexical density, diversity and sophistication measures. Subsequently, all of the measures were tested for multicollinearity on SPSS26 and the remainders of lexical complexity measures were retained and presented in Table 2.

Table 2. Lexical density, lexical diversity and lexical sophistication indices

No.	Index	Description	Dimension
1.	lexical_density_types lexical_density_tokens	Content word types (N) divided by word types (N)	Lexical density

		Content word tokens divided by word tokens	
2.	root_ttr_aw root_ttr_cw root_ttr_fw	Guiraud for all word types Guiraud for content words Guiraud for function words	Lexical diversity (Guiraud)
3.	hdd42_aw hdd42_cw hdd42_fw	Hypergeometric Distribution D for all words Hypergeometric Distribution D for content words Hypergeometric Distribution D for function words	Lexical diversity
4.	mtld_original_aw mtld_original_cw mtld_original_fw	Average number of all tokens required to reach TTR >= .720 Average number of all content word tokens to reach TTR >= .720 Average number of all function word tokens to reach TTR >= .720	Lexical diversity
5.	BNC_Written_Freq_AW_Log BNC_Written_Freq_CW_Log BNC_Written_Freq_FW_Log BNC_Written_Range_AW BNC_Written_Range_CW BNC_Written_Range_FW	Average log frequency of all words in BNC written corpus Average log frequency of content words the BNC written corpus Average log frequency of function words in the BNC written corpus Breadth of all word distribution across texts Breadth of content word distribution across texts Breadth of function word distribution across texts	Lexical sophistication *****Higher values = more common, basi**c** vocabulary overall
6.	All_AWL_Normed	Academic Word List All	Lexical sophistication
7.	aoe_inverse_average	LDA Age of Exposure (inverse average)	Lexical sophistication

3.2.2. Syntactic complexity measures

Similar to lexical complexity, 14 holistic syntactic complexity measures were derived from the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASC; Kyle, 2016) and presented in Table 3.

Table 3. Holistic syntactic complexity measures

No.	Index	Description	Category
1.	MLS MLT MLC	Mean length of sentence Mean length of T-unit Mean length of clause	Length-based indices (3)
2.	C/S	Clauses per sentence	Sentence complexity (1)
3.	C/T CT/T DC/C DC/T	Clauses per T-unit Complex T-units per T-unit Dependent clauses per clause Dependent clauses per T-unit	Subordination indices (4)
4.	CP/C CP/T	Coordinate phrases per clause Coordinate phrases per T-unit	Coordination (3)

	T/S	T-units per sentence	
5.	VP/T	Verb phrases per T-unit	Phrasal indices (3)
	CN/C	Complex nominals per clause	
	CN/T	Complex nominals per T-unit	

3.2.2.1. Phrasal complexity

Previous research highlights the significance of phrasal complexity measures related to complex noun phrases, nominals, embedding information in noun phrases, etc. as critical determiners of writing proficiency (Kyle, 2016; Crossley & McNamara, 2014; Guo et al., 2013; McNamara et al., 2010). In TAASSC, measures of phrasal complexity are based on seven noun phrase types (e.g., nominal subject) and ten phrasal dependent types (e.g., adjectival modifiers). In this study, 15 indices of phrasal complexity were chosen, following Nguyen and Le (2024)'s framework of assessing reading text complexity.

Table 4. Phrasal complexity measures

No.	Index	Description
1.	av_nsubj_deps	Dependents per nominal subject
2.	av_ncomp_deps	Dependents per nominal complement
3.	av_dobj_deps	Dependents per direct object
4.	av_iobj_deps	Dependents per indirect object
5.	av_pobj_deps	Dependents per prepositional object
6.	det_all_nominal	Determiners per nominal phrases
7.	amod_all_nominal	Adjective modifiers per nominal phrases
8.	prep_all_nominal	Prepositional phrases per nominal phrases
9.	poss_all_nominal	Possessives per nominal phrases
10.	vmod_all_nominal	Verbal modifiers per nominal phrases
11.	nn_all_nominal	Nouns as modifiers per nominal phrases
12.	rcmod_all_nominal	Relative clause modifiers per nominal phrases
13.	advmod_all_nominal	Adverbial modifiers per nominal phrases
14.	conj_and_all_nominal	Conjunctions "and" per nominal phrases
15.	conj_or_all_nominal	Conjunctions "or" per nominal phrases

3.2.2.2. Syntactic sophistication

As previously reviewed, measures of syntactic sophistication indicate the relative complexity of the syntactic structures used in production. In this study, several measures by frequency and association strength were chosen from TAASSC as in Table 5.

Table 5. Syntactic sophistication indices chosen from TAASSC

Academic VAC Frequency Measures	Academic VAC Strength Measures
acad_av_lemma_freq	acad_av_approx_collexeme
acad_av_construction_freq	acad_av_faith_verb_cue

acad_av_lemma_construction_freq	acad_av_faith_const_cue
acad_av_lemma_freq_log	acad_av_delta_p_verb_cue
acad_av_construction_freq_log	acad_av_delta_p_const_cue
acad_av_lemma_construction_freq_log	acad_collexeme_ratio

The multicollinearity tests on SPSS 26 for the chosen variables revealed that acad_av_construction_freq was found to significantly predict word count (β = .336, p = .044), indicating that constructions are associated with longer utterances. Additionally, acad_av_lemma_freq_log showed a marginal negative effect (β = –.286, p = .059), suggesting that more frequent lemmas might be linked to shorter outputs. Other variables did not show statistically significant contributions, though acad_av_faith_const_cue approached concerning VIF levels (4.4), indicating potential multicollinearity.

Therefore, the finalized measures for Academic VAC Frequency include: acad_av_construction_freq, acad_av_construction_freq_log and Academic VAC strength indices include: acad_av_approx_collexeme and acad_collexeme_ratio.

3.2.3. Discoursal complexity:

In this study, cohesion measures were selected from TAACO. Based on the nature of their linguistic features, cohesive indices were classified into four functional categories, i.e. lexical overlaps, semantic overlaps, connectives and givenness as presented in Table 6.

Table 6. Cohesion measures from TAACO

No.	Functional Category	Cohesive Index	Description
1	Lexical Overlap	adjacent_overlap_noun_sent	Noun repetition across adjacent sentences
2	Lexical Overlap	adjacent_overlap_binary_2_noun_para	Binary indicator of noun overlap across adjacent paragraphs
3	Lexical Overlap	adjacent_overlap_2_verb_para	Verb repetition across adjacent paragraphs
4	Lexical Overlap	syn_overlap_sent_noun	Syntactic overlap of nouns between adjacent sentences
5	Lexical Overlap	syn_overlap_sent_verb	Syntactic overlap of verbs between adjacent sentences
6	Lexical Overlap	syn_overlap_para_noun	Noun overlap across paragraphs
7	Lexical Overlap	syn_overlap_para_verb	Verb overlap across paragraphs
8	Semantic Overlap	lsa_2_all_para	Latent Semantic Analysis (LSA) similarity across paragraphs
9	Semantic Overlap	lda_2_all_para	Topic modeling (LDA) similarity across paragraphs
10	Semantic Overlap	word2vec_1_all_para	Semantic similarity between adjacent paragraphs
11	Semantic Overlap	word2vec_2_all_para	Semantic similarity between all paragraph pairs
12	Semantic Overlap	lsa_1_all_sent	LSA-based semantic similarity across all sentences in the text

13	Semantic Overlap	lda_1_all_sent	LDA-based topic similarity across all sentences
14	Semantic Overlap	word2vec_1_all_sent	Word2Vec-based semantic similarity across all sentences
15	Semantic Overlap	word2vec_2_all_sent	Broader context semantic similarity between sentences
16	Connectives	basic_connectives	Use of basic conjunctions (e.g., and, but, so) to connect clauses/sentences
17	Connectives	sentence_linking	Explicit linking of sentences using logical operators or connectors
18	Connectives	all_additive	Use of additive connectives like also, moreover
19	Connectives	all_causal	Use of causal connectives like because, since
20	Connectives	all_temporal	Use of temporal connectives like then, finally
21	Connectives	reason_and_purpose	Logical purpose-based connectives used across sections
22	Connectives	all_logical	Logical relations like however, therefore connecting larger text units
23	Givenness	attended_demonstratives	Demonstratives with clear referents (e.g., this idea)
24	Givenness	unattended_demonstratives	Demonstratives without explicit referents (e.g., this)
25	Givenness	pronoun_density	Frequency of pronouns linking back to earlier content
26	Givenness	pronoun_noun_ratio	Ratio of pronouns to nouns, reflecting cohesive referencing
27	Givenness	repeated_content_lemmas	Repetition of content words (lemmas) across nearby text
28	Givenness	repeated_content_and_pronoun_lemmas	Repetition of content words and pronouns
29	Givenness	all_demonstratives	All types of demonstratives contributing to reference chains

3.3. Data analysis

All of the selected measures of linguistic complexity across lexical, syntactic and textual domains were fed into SPSS 26. Shapiro-Wilk tests were performed to test for normality. Since the dataset did not have normal distribution, Kruskal Wallis tests were employed to find out significant differences between each of the complexity indices across the five learners. If statistically significant results were found, post-hoc analysis was conducted by running Mann-Whitney U Tests. To control for Type I error due to multiple comparisons, a Bonferroni correction was applied. Given 10 pairwise tests (since there were 5 learners), the adjusted significance threshold was set at $\mathbf{p < 0.005}$. The effect size (r) was calculated using the formula $r = Z / \sqrt{N}$, where N is the total number of observations for each pairwise comparison.

To test the inter-rater reliability, both Cronbach's Alpha and the Intraclass Correlation Coefficient (ICC) were assessed. Cronbach's Alpha yielded was 0.7. Inter-rater reliability was calculated using a two-way mixed effects model with absolute agreement. The Intraclass Correlation Coefficient (ICC) for single measures was 0.500 (95% CI: 0.257–0.680), and that for average measures was 0.667 (95% CI: 0.409–0.809). The F-test result was statistically significant ($F(55, 55)$

= 3.338, $p < .001$). Since the overall average score of Judge 2 failed to have normal distribution, Spearman Rho was run to confirm the inter-rater reliability (Spearman's ρ =0598, p=0). The aggregate results reflected the moderate consistency of the two raters in scoring the IELTS essays.

4. RESULTS AND DISCUSSION

4.1. Results

Research question 1: Differences in lexical complexity across 5 learners

There were statistically significant differences in lexical diversity (Guiraud, HDD, MTLD) and vocabulary range (BNC-based measures). Five pairwise comparisons between learners showed statistically significant differences ($p < .005$, Bonferroni corrected) in lexical complexity indices. All yielded large effect sizes ($r > .5$), indicating meaningful differences in lexical diversity and vocabulary range among individual learners (see Table 8).

Table 7. Significant differences in lexical complexity across five learners

Variable	H (Kruskal-Wallis)	p-value
root_ttr_aw	11.781	0.019
root_ttr_fw	9.492	0.050 (baseline)
hdd42_aw	14.318	0.006
mtld_original_aw	9.678	0.046
mtld_original_fw	12.342	0.015
BNC_Written_Range_CW	9.887	0.042

Table 8. Post-hoc & effect size (Mann-Whitney U test)

Group Comparison	Variable	U-value	p-value	Z-value	Effect Size (r)
S1 vs S2	mtld_original_aw	12.000	0.003	-2.849	0.637
S1 vs S4	mtld_original_fw	19.000	0.002	-3.042	0.621
S1 vs S5	BNC_Written_Range_CW	19.000	0.005	-2.725	0.581
S2 vs S3	root_ttr_aw	15.000	0.004	-2.773	0.605
S2 vs S3	hdd42_aw	1.000	0.000	-3.767	0.823

Learner 3 demonstrated a higher root TTR and HD-D (all words) than Student 2, indicating greater lexical diversity in Student 3's writing. Student 1 produced texts with significantly more lexical variation compared to Student 2, suggesting a richer and more varied vocabulary (MTLD (all words)) whereas Student 4 used a wider range of function words than Student 1(MTLD (function words)), which may reflect a more complex grammatical structure. Student 5 used content words with broader range (more widely distributed across BNC written texts) than Student 1, possibly suggesting more conventional word choice.

Research question 2: Differences in syntactic complexity across 5 learners

Table 9. Significant syntactic complexity indices

Variable	H (Kruskal-Wallis)	p-value
T_S	10.386	0.034
CP_C	11.603	0.021

The Kruskal-Wallis H test revealed statistically significant differences across the five learners in two syntactic complexity indices: T_S (T-units per sentence) and CP_C (co-ordinate phrases per clause), with p-values of 0.034 and 0.021 respectively.

Table 10. Post-hoc & effect size (Mann-Whitney U test)

Group Comparison	Variable	U-value	p-value	Z-value	Effect Size (r)
S1 vs S2	T_S	11.500	0.002	-3.028	0.667
S2 vs S3	CP_C	12.500	0.002	-2.950	0.644

After applying the Bonferroni correction for multiple pairwise comparisons (adjusted $\alpha = 0.005$), two comparisons remained statistically significant: (1) Learner 2 wrote significantly more T-units per sentence (T_S) than Learner 1 (U = 11.500, p = 0.002, r = 0.677, large effect), indicating a higher level of sentence complexity. (2) Learner 3 produced significantly more co-ordinate phrases per clause (CP_C) than Learner 2 (U = 12.500, p = 0.002, r = 0.644, large effect), suggesting stronger syntactic subordination skills. These results highlight meaningful differences in syntactic development between individual learners.

Research question 3: Significant differences in discoursal complexity across 5 learners

Table 11. Significant discoursal complexity indices

Cohesion Index	Kruskal-Wallis H	p-value
syn_overlap_sent_verb	10.834	0.028
word2vec_1_all_sent	21.129	0.000
all_causal	11.069	0.026
all_additive	10.660	0.031
repeated_content_lemmas	14.774	0.005
repeated_content_and_pronoun_lemmas	19.400	0.001

The Kruskal-Wallis test identified statistically significant differences across learners in six cohesion indices. These include: lexical cohesion (verb) at the sentence level (syn_overlap_sent_verb), semantic similarity between sentences (word2vec_1_all_sent), use of causal and additive connectives (all_causal, all_additive), givenness-based cohesion, reflected in both repeated content lemmas and repeated content with pronouns.

Table 12. Post-hoc & effect size (Mann-Whitney U test)

Group Comparison	Cohesion Index	U-value	Z-value	p-value	Effect Size (r)
2 vs 3	all_additive	14.000	-2.843	0.003	0.620
1 vs 5	all_causal	18.000	-2.791	0.004	0.595
1 vs 3	repeated_content_and_pronoun_lemmas	5.000	-3.754	0.000	0.783
1 vs 5	repeated_content_and_pronoun_lemmas	11.000	-3.250	0.001	0.693
3 vs 4	repeated_content_and_pronoun_lemmas	27.000	-2.774	0.005	0.555
1 vs 3	repeated_content_lemmas	7.000	-3.631	0.000	0.757
3 vs 4	repeated_content_lemmas	11.000	-3.645	0.000	0.729
1 vs 2	word2vec_1_all_sent	8.000	-3.153	0.001	0.705
1 vs 4	word2vec_1_all_sent	15.000	-3.273	0.001	0.668
2 vs 3	word2vec_1_all_sent	13.000	-2.914	0.002	0.636
3 vs 4	word2vec_1_all_sent	28.000	-2.720	0.005	0.544

The post-hoc analysis using the Mann-Whitney U test revealed several statistically significant differences ($p < .005$, Bonferroni-corrected) in cohesion indices across learner pairs with large effect sizes ($r > .5$). word2vec_1_all_sent consistently showed significant differences between multiple learner pairs (e.g., 1 vs 2, 1 vs 4, 2 vs 3, 3 vs 4), with r ranging from 0.603 to 0.738. This suggests that learners differ substantially in their ability to maintain semantic relatedness across adjacent sentences, a key aspect of textual cohesion and reader comprehension. Givenness-based indices **(**repeated_content_lemmas and repeated_content_and_pronoun_lemmas) appeared repeatedly with significant results, particularly between Learner 1 and Learner 3 or 4. High variation here implies differing levels using lexical repetition and pronominal reference. Learner 1 differed significantly from Learner 5 in the use of causal connectives, and Learner 2 differed from Learner 3

in the use of additive connectives, revealing the divergence in how learners structure argumentation or explanation.

Research question 4: Compatibility between variation in linguistic complexity measures and human judgement

The ANOVA results indicate significant differences in human judgments of essay quality among the five learners (Table 12). Learner 4 received the highest scores, outperforming all other participants. Learner 3 followed, scoring lower than Learner 4 but higher than Learners 1 and 5. There was no significant difference between Learner 3 and Learner 2, though Learner 2 performed better than Learner 5 but worse than Learner 4. Learner 1 and Learner 2 did not differ significantly in their essay scores; however, Learner 1 scored lower than Learners 3 and 4, yet higher than Learner 5. Consistently, Learner 5 obtained the lowest scores among the group. Overall, the ranking of essay quality based on human ratings, from highest to lowest, is as follows: Learner 4, Learner 3, Learner 2, Learner 1, and Learner 5.

Table 13. Comparison between complexity variation and holistic human ratings

Pairs of Learners	Mean difference	p-value	Complexity measures with significant variation	p-value
Learner 1 – Learner 2	–0.28384	0.120	mtld_original_aw T_S word2vec_1_all_sent	0.003 0.002 0.001
Learner 1 – Learner 3	**–0.43106***	**0.003**	repeated_content_and_pronoun_lemmas repeated_content_lemmas	0.000 0.000
Learner 1 – Learner 4	**–0.85734***	**0.000**	mtld_original_fw word2vec_1_all_sent	0.002 0.001
Learner 1 – Learner 5	**+0.35455***	**0.035**	BNC_Written_Range_CW all_causal repeated_content_and_pronoun_lemmas	0.005 0.004 0.001
Learner 2 – Learner 3	–0.14722	0.727	root_ttr_aw hdd42_aw CP_C all_additive word2vec_1_all_sent	0.004 0.000 0.002 0.003 0.002
Learner 2 – Learner 4	**–0.57350***	**0.002**		
Learner 2 – Learner 5	**+0.63838***	**0.001**		
Learner 3 – Learner 4	**–0.42628***	**0.016**	repeated_content_and_pronoun_lemmas repeated_content_lemmas word2vec_1_all_sent	0.005 0.000 0.000
Learner 3 – Learner 5	**+0.78561***	**0.000**		
Learner 4 – Learner 5	**+1.21189***	**0.000**		

The analysis reveals that in some learner pairs, significant differences in essay scores are associated with significant variations in linguistic complexity measures, while in other pairs, differences in scores do not coincide with notable variations in complexity. Specifically, for the pair Learner 1 – Learner 3**,** there is a significant difference in essay scores ($p = 0.003$), which corresponds with significant differences in repeated content and pronoun lemmas ($p = 0.000$). Similarly, between Learner 1 – Learner 4, the essay score difference is highly significant ($p = 0.000$), alongside significant variation in lexical diversity (mtld_original_fw, $p = 0.002$) and semantic cohesion (word2vec_1_all_sent, $p = 0.001$). However, for pairs like Learner 2 – Learner 4, Learner 2 – Learner 5, Learner 3-Learner 5, Learner 4-Learner 5 although there are significant differences in essay scores ($p = 0.002$, $p = 0.001$, $p = 0.000$, $p = 0.000$ respectively), no complexity measures show significant variation. This means that differences in human-rated essay quality may not be not fully explained by variations in linguistic complexity and other factors such as accuracy and content relevance.

Conversely, for Learner 1 – Learner 2, Learner 2 – Learner 3, although there is no significant difference in essay scores, significant variations exist in lexical and semantic measures ($p < 0.01$).

4.2. Discussion

This study was built upon the usage-based perspective with an assumption that language use and production are inherently individualized processes, shaped by each learner's unique trajectory of internalization (Langacker, 2002) and that even under similar instructional conditions, learners process, practice, and produce language differently, resulting in distinct patterns of linguistic complexity.

Research question 1:
There were differences in the use of lexical complexity, specifically the lexical diversity and lexical sophistication. The meaningful inter-learner variation was found between multiple pairs of learners, indicating the individual differences are of complex nature.

Although lexical density did not differ among learners, their writings showed variation in lexical diversity and sophistication. This aligns with the absence of differences in content words (CW), but notable differences in all words (AW) and function words (FW). The similarity in CW and in age of acquisition index (aoe_inverse_average) suggests learners shared comparable topic-related vocabulary, likely due to similar instructional input or comparable exposure to academic English.

In contrast, significant differences in root_ttr_aw, mtld_aw, and mtld_fw indicate disparities in syntactic and discourse management. Since FW structure sentences and maintain cohesion, greater FW variation in one learner points to more flexible and sophisticated language use. This also explains AW differences, as AW reflects combined variation from CW and FW. While learners expressed similar content, their use of structure and transitions differed.

Furthermore, although hdd42 shows no differences in content words and functional words, the difference in all words suggests that subtle shifts in how words are distributed may account for variation, highlighting the need to assess both overall and category-specific lexical diversity (McCarthy & Jarvis, 2010).

Of the measures selected for lexical sophistication, only BNC_range_cw showed differences between learners whereas they used words of similar overall frequency (BNC_Frequency_Log) and similar age of acquisition index (aoe_inverse_average). This result suggests variation in how broadly their vocabulary was distributed across discourse types.

To summarize, there exists inter-learner variability within the domain of lexical complexity, except for lexical density. These findings well resonate the assumptions of usage-based theory in second language acquisition/learning in that L2 production is affected through usage such as learning strategies, engagement, experience (Verspoor et al., 2012, 2017).

Research question 2:

No significant differences were found between learners in fine-grained phrasal complexity indices, such as complex nominals per clause or noun phrase modifiers. This result may suggest that both learners operate at a similar level of syntactic development in academic writing, particularly in the use of elaborated noun phrases. The absence of variation may reflect shared instructional input or similar exposure to academic writing norms. Another explanation is that because fine-grained indices are more stable and captures highly specific syntactic features that occur at relatively low frequencies in short to mid-length L2 texts, subtle differences between learners may not be easily detected without larger sample sizes or more cognitively demanding tasks (Biber et al., 2011; Lu, 2011).

Meaningful inter-learner variation in syntactic complexity was found. First, Learner 2 produced significantly more T-units per sentence (T_S) than Learner 1, indicating a higher level of sentence-level complexity. Learner 2's writing may exhibit more complex sentence structures with greater syntactic depth, demonstrating stronger control over clause linkage and rhetorical organization, both of which contribute to advanced academic writing proficiency. Second, Learner 3 produced more coordinate phrases per clause (CP_C) than Learner 2, reflecting greater use of coordination strategies within clauses. While coordination is generally considered a simpler form of syntactic expansion compared to subordination, a high frequency of well-controlled coordinate phrases can also reflect differences in writing style (Norris & Ortega, 2009).

Research question 3:

Findings reveal a wide range of complex individual variation in using discoursal complexity. The six indices that showed significant differences reflect multiple dimensions of textual cohesion, ranging from lexical overlaps, to semantic overlaps, and information flow management (givenness). This suggests that learners vary not only in lexical repetition but also in how they structure meaning and maintain discourse continuity across texts.

The findings revealed large and consistent inter-learner differences across multiple cohesion indices, particularly in additive and causal connectives, referential repetition, and semantic similarity, as measured by *word2vec* embeddings. These differences suggest that learners employ distinct strategies for maintaining cohesion, spanning from surface-level lexical repetition to deeper semantic and discourse-level organization.

Learner 2 and 3 differed significantly in their use of additive connectives (e.g., *and*, *also*), while Learner 1 and 5 showed strong contrasts in the use of causal connectives (e.g., *because*, *therefore*). These findings echo Crossley et al. (2016), who found that more proficient L2 writers tend to use a wider range of logical connectors to guide readers through the argument structure, thereby enhancing global cohesion.

Furthermore, repeated content lemmas and pronouns, particularly between Learner 1 and Learners 3, 4, and 5, demonstrate variability in textual cohesion. Learners who frequently recycle core content words and referents (e.g., through pronouns or lexical repetition) are better able to signal topic maintenance and reinforce thematic focus, consistent with findings by Halliday and Hasan (1976) and more recent work by Crossley et al. (2016).

Significant differences also emerged in semantic similarity across sentences using word2vec-based indices. Learners who consistently link sentences semantically (e.g., Learner 3 vs. 4, 1 vs. 2) demonstrate stronger conceptual integration, suggesting more coherent overall discourse. As shown in Crossley et al. (2016), such semantic overlaps are associated with higher text quality and reader comprehension.

Research question 4:

According to usage-based linguistics, linguistic structures emerge from repeated exposure to, and practice with, meaningful language use. Complexity is not a fixed trait but develops gradually through frequency, salience, and entrenchment (Verspoor et al., 2012). This aligns with the observation that some learners, such as Learner 1 vs. Learner 3 and Learner 1 vs. Learner 4, show both significant score differences and significant variations in usage-driven measures, such as: lexical diversity (MTLD), semantic cohesion (word2vec_1_all_sent), indicating more entrenched patterns of meaning-making across discourse, givenness (repeated content/pronoun lemmas).

However, variation in complexity measures does not uniformly predict differences in human ratings. In pairs such as Learner 2 vs. Learner 4, Learner 2-Learner 5, Learner 3-Learner 5, Learner 4-Learner 5 despite a significant difference in scores, no significant complexity variations were observed. This reflects a core critique of complexity indices raised by Norris & Ortega (2012) and Bulté & Housen (2014): measures often capture surface-level variation and fail to account for accuracy, appropriateness, or content richness. This suggests that holistic performance encompasses more than complexity alone. Other dimensions like fluency, error rates, task relevance, and discourse appropriateness may have driven raters' judgments in these cases.

Interestingly, in pairs such as Learner 1 – Learner 2 and Learner 2 – Learner 3, significant differences in complexity measures (e.g., MTLD, semantic cohesion) emerged without significant score differences. From a usage-based perspective, this may reflect emerging but not yet fully functional complexity. Learners may experiment with complex forms (e.g., more diverse vocabulary, semantic elaboration) without successfully integrating them into communicatively effective performance. This aligns with Skehan's (2009) view that trade-offs between complexity, accuracy, and fluency can result in learners prioritizing form over function.

These findings reinforce the notion that complexity is a dynamic, multidimensional construct. In some cases, it correlates with perceived proficiency, particularly when it comes to semantic cohesion and lexical control. In others, complexity varies independently of performance, suggesting that the developmental stage of complexity or task-related factors (e.g., topic familiarity, content coherence) may moderate its impact on ratings. The non-linear relationship observed here supports the dynamic model of L2 development (Verspoor, Lowie & de Bot, 2008), where variation and emergence are expected, and performance is seen as adaptive rather than static.

5. CONCLUSION

Learners showed individual variation in their use of linguistic complexity across lexical, syntactic, and discourse levels. Significant differences between several learner pairs suggest that these individual differences are intricate and multifaceted in nature. From a usage-based perspective, the relationship between linguistic complexity and human-rated performance is context-sensitive and shaped by the learner's accumulated linguistic experience. While complexity measures can predict performance in certain contexts, especially when reflecting entrenched, semantically cohesive usage, they cannot fully explain score differences across all learner pairs. These findings emphasize the need for personalized instruction since learners develop linguistic complexity in diverse and individual ways. Teachers should emphasize functional use of complexity to enhance communication rather than making the discourse sound more complex. Additionally, providing learners with rich input and meaningful practice can support the emergence of cohesive and sophisticated language use over time.

DISCLOSURE STATEMENT

The author acknowledges the use of ChatGPT, an AI language model developed by OpenAI, to assist in refining the academic style and language clarity of this manuscript. The content, analysis, and interpretations presented remain entirely the author's original work.

REFERENCES

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Bui, H. P. (2022). Vietnamese EFL Students' Use and Misconceptions of Cohesive Devices in Writing. *Sage Open*, *12*(3), 21582440221126993. https://doi.org/10.1177/21582440221126993

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

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Kuiken, F. (2023). Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), 83–93. https://doi.org/10.1515/lingvan-2021-0112

Kuiken, F., Vedder, I., Housen, A., & De Clercq, B. (2019). Variation in syntactic complexity: Introduction. *International Journal of Applied Linguistics*, *29*(2), 161–170. https://doi.org/10.1111/ijal.12255

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication* [Georgia State University]. https://doi.org/10.57709/8501051

Langacker, R. W. (2002). Concept, image, and symbol: The cognitive basis of grammar (2nd ed.). Mouton de Gruyter. https://doi.org/10.1515/9783110874014.

Lu, X. (2025). Meaning and function dimensions of linguistic complexity in second language writing. *Research Methods in Applied Linguistics*, *4*(1), 100191. https://doi.org/10.1016/j.rmal.2025.100191

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.

O'Leary, J. A., & Steinkrauss, R. (2022). Syntactic and lexical complexity in L2 English academic writing: Development and competition. *Ampersand*, *9*, 100096. https://doi.org/10.1016/j.amper.2022.100096

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Nguyen, H. T. M., & Le, N. V. A. (2024). Text Complexity of Cambridge-delivered IELTS Academic Reading Tests: Comparability with IELTS Academic Reading Practice Tests from Other Publishers. *Teaching English as a Second or Foreign Language--TESL-EJ*, *28*(2). https://doi.org/10.55593/ej.28110a4

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Verspoor, M., Lowie, W., Chan, H. P., & Vahtrick, L. (2017). Linguistic complexity in second language development: Variability and variation at advanced stages. *Recherches En Didactique Des Langues et Des Cultures*, *14*(1). https://doi.org/10.4000/rdlc.1450

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239–263. https://doi.org/10.1016/j.jslw.2012.03.007

Verspoor, M., Lowie, W., & de Bot, K. (2008). Dynamic systems theory and variation: A case study in L2-writing. *Language Learning*, *58*(2), 243–283. https://doi.org/10.1111/j.1467-9922.2008.00490.x

Zhang, J., & Zhang, L. J. (2023). Lexical cohesion development in English as a foreign language learners' argumentative writing: A latent class growth model approach. *Linguistics and Education*, *78*, 101255. https://doi.org/10.1016/j.linged.2023.101255