ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 15, No. 3, pp. 772-786, May 2024
DOI: https://doi.org/10.17507/jltr.1503.10

A Corpus-Based Study on Cohesion Development in Writing by EFL Novices

Weilu Wang
Foreign Languages College, Inner Mongolia University, Hohhot, China

Wei Chen
Foreign Languages College, Inner Mongolia University, Hohhot, China

Shuangshuang Shi
Foreign Languages College, Inner Mongolia University, Hohhot, China

Lin Tong
Foreign Languages College, Inner Mongolia University, Hohhot, China

Manfu Duan[*]
Division of International Cooperation and Exchange, Inner Mongolia University, Hohhot, China

Abstract—**This corpus-based study explored the development of cohesive devices in the writing of Chinese beginner learners of English as a foreign language (EFL) over a three-year span. Quantitative analysis utilizing the Tool for Automatic Analysis of Cohesion (TAACO) was conducted on a longitudinal learner corpus comprising over 500 exam essays. Lexical, syntactic, semantic, and discourse features were examined to identify reliable indices for tracking learners' progressive mastery of cohesion. Results revealed that pronoun-related features, including pronoun density and repetition, significantly differed across year pairs and robustly predicted writing development. However, most lexical and connective indices showed ambiguous trajectories over time. The findings highlight the vital role of pronouns in building coherence for novice writers and underscore persistent difficulties in acquiring sophisticated content words and their collocations. This study contributes data-driven insights into the nonlinear processes and enduring challenges shaping EFL beginners' cohesive competence. It demonstrates the value of computational tools and learner corpora in exploring discourse acquisition.**

Index Terms—**learner corpus, cohesion development, English writing, EFL novice**

I. Introduction

In Second Language Acquisition (SLA), cohesion has long been recognized as a crucial element in constructing meaning and conveying ideas effectively. Cohesion refers to the links that hold a text together and allow it to be interpreted as a meaningful whole by signaling the relationship between sentences and paragraphs (Halliday & Hasan, 1976). In recent years, the significance of cohesion in the writing of EFL learners has been widely emphasized. For second language learners, using appropriate and effective cohesive devices poses a significant challenge and is an essential indicator of their writing ability and language development. Mastering cohesion is closely related to learners' proficiency and discourse competence (Wray & Perkins, 2000). Studies have shown that lack of cohesion is a distinguishing feature of learner writing, and cohesion use differentiates high-proficiency learners from low-proficiency ones (Connor, 1984; Zhang, 2000).

In a broader sense, cohesion is a central component of discourse competence and is closely linked to an individual's language proficiency and cognitive development (Wray & Perkins, 2000). Analyzing cohesive patterns in learner language allows researchers to better understand learners' abilities in connecting ideas and navigating discourse (Granger & Tyson, 1996), as well as to trace their language learning progress over time. Previous studies have shown the significant role of cohesive devices in predicting writing quality and differentiating developmental stages (Witte & Faigley, 1981; Crossley & McNamara, 2011).

On the frontier, learner corpus research has become increasingly prevalent in second language acquisition in recent decades, including the study of cohesion in writing. By amassing and analyzing learners' authentic language data, learner corpus studies have provided valuable insights into learner language that complement traditional qualitative studies (Granger, 2002). Compared with natural language corpora, the "ecological validity" of learner corpora allows researchers to observe genuine learning processes and outcomes (Meunier, 2002, p. 138). The use of computational

[*] Corresponding Author. E-mail: duanmanfu@imu.edu.cn

techniques also enables large-scale analyses of linguistic features in learner language. These advantages have led to a growing interest in building learner corpora and a surge in studies examining various aspects of learner language. In recent years, NLP techniques have been used to investigate cohesion in learner language. For example, automated cohesion analyses can identify and quantify cohesive devices in learners' writing, such as lexical repetition (Crossley & McNamara, 2011), grammatical cohesion, and conjunctions (Kong & Pearson, 2003). By comparing such cohesive profiles of learners at different proficiency levels, NLP tools provide a means to explore developmental patterns in acquiring cohesive competence. For example, Crossley and McNamara (2011) found significant differences in lexical bundle and grammatical cohesion between high- and low-rated TOEFL essays. Such large-scale analyses enabled by NLP would be impossible through manual annotation.

This study investigates cohesion in English writing texts by EFL beginners using a learner corpus approach. By analyzing the frequency and use of a set of cohesive devices with the help of an NLP tool named TAACO. This study aims to explore the predictability of a set of statistical indices to determine the ones that count in marking the development of cohesion in beginner writing. The findings provide valuable insights into the cohesion learning process of EFL beginners and their potential difficulties. They can also inform EFL writing instruction by helping instructors identify key areas to focus on in guiding students toward producing more cohesive writing over time. In summary, this study seeks to contribute to the growing field of learner corpus research by exploring cohesion use in EFL beginners' writing and providing implications for writing instruction and automated writing evaluation. The findings will also enrich our understanding of the language learning process from a discourse perspective. This study aims to yield more comprehensive and ecologically valid insights into EFL beginners' cohesion learning and inform pedagogical implications for their writing development.

II. Literature Review

A. Cohesion of Learner English

Previous research has established that cohesion is one of the critical features of effective second-language writing (Bitchener & Basturkmen, 2006; Hyland, 2004). Cohesion refers to the explicit and implicit connections among words, phrases, and sentences in texts, which are crucial for conveying meaning and coherence. Many studies have investigated the use of cohesive devices by EFL learners to identify areas of difficulty. Some specific challenges that EFL learners face in using cohesive devices include the correct use of pronouns, conjunctions, tense and aspect markings, prepositions, and articles (Tsou, 2005; Hyland, 2004). Few studies have specifically used corpus-based approaches to examine cohesion in beginner-level EFL writing. Moreover, large-scale investigations of global cohesion in beginner writing are still scarce.

B. NLP for Cohesion

Recent research has also used various NLP techniques and statistical measures to assess beginner writing quality based on cohesion. For example, McNamara et al. (2013) identified lexical bundles, grammatical cohesion, and word frequency indices that can distinguish high- and low-rated beginner essays. Crossley et al. (2016) developed automated indices of lexical sophistication, syntactic complexity, and cohesion (e.g., lexical repetition) to predict essay scores. Shin and Kim (2017) proposed coherence indices based on entity grid, centering, and lexical chain to measure coherence in L2 writing. Their studies prove the feasibility of assessing beginner writing through computational methods. These innovative studies have demonstrated the significant potential of using NLP and statistical techniques to gain a deeper understanding of challenges faced by beginner EFL writers, especially in constructing coherent discourse. NLP and statistical techniques have assessed beginner writing quality based on coherence and cohesion. Studies have identified various linguistic indices, developed predictive models, and designed automated evaluators using these techniques. The computational methods complement traditional qualitative evaluations by enabling large-scale analyses and pinpointing areas of difficulty for beginners. They provide promising means for tracing learners' progress over time and supporting their writing development through data-driven insights and feedback.

However, fewer studies were conducted specifically on beginners from a longitudinal perspective. To address this gap, the current study aims to conduct a comprehensive analysis of cohesive features used by beginner-level EFL writers via corpus-based methods and NLP indices. By exploring the development of cohesion in learner writing corpus, this research aims to provide insights into the challenges faced by beginner-level EFL writers in constructing coherent texts. The findings may have significant implications for language teaching and learning, as they can inform instructional strategies that target the identified areas of difficulty for learners at this proficiency level. To serve the purpose, the following research questions are addressed:

1. How does cohesion in EFL beginners' writing develop over the years?
2. What cohesion indices are most useful in determining beginner writing development?
3. What difficulties beginners may face in achieving cohesion in their writing?

III. Method

A. Corpus Compilation

The learner corpus utilized in this study comprises English essays compiled from an intensive English program at a Chinese university over a 3-year span (2017-2020). The essays were collected from approximately 170 freshman students enrolled in the program who represent beginner-level EFL writers. The students attend four hours of weekly English classes focusing on core language skills including basic writing. The essays were drawn from timed in-class exams administered at the conclusion of each academic term, ensuring the texts exemplify students' authentic writing abilities. The essay prompts spanned a range of topics such as technology, education, environment, and lifestyle.

To construct the electronic corpus, the handwritten essays were digitized via Optical Character Recognition (OCR) techniques and manually checked for accuracy against the original texts. Any errors in conversion were manually corrected. The final corpus contains over 500 essays (70,000 tokens). For analysis, the corpus was divided into three sub- corpora based on students' year of study - first, second, and third year. This enables comparisons of developmental trajectories over the years.

B. Cohesion Analysis

To conduct automated cohesion analysis, this study utilized the Tool for the Automatic Analysis of Cohesion (TAACO) version 2.0.4. TAACO calculates 150 validated indices of local, global, and overall textual cohesion encompassing diverse dimensions including lexical diversity, semantic overlap, connectives, givenness, and lexical repetition across sentences and paragraphs (Crossley et al., 2016). The comprehensive indices allow the examination of multiple facets of cohesive device usage in learner writing.

Specifically, the TAACO indices were applied to each essay in the corpus, and computed scores were compared statistically across sub-corpora grouped by year of study. Due to non-normal distribution, the non-parametric Kruskal-Wallis H test was utilized to determine significant differences among year groups for each index. Indices demonstrating statistically significant differences for multiple comparisons were interpreted as potentially reliable markers of developmental progression in novices' cohesive competence. Through this quantitative analysis of patterns in TAACO index scores over time, this study aimed to pinpoint specific dimensions of cohesion that pose persistent challenges for beginner writers versus areas of growth.

The automated TAACO measures complement traditional qualitative methods by enabling robust large-scale analysis of cohesive features and precise quantification of linguistic phenomena. The findings provide data-driven insights into developmental shifts in beginners' ability to construct coherent discourse and achieve textual unity. Moreover, the results inform future instructional interventions targeting particular facets of cohesion that students struggle to master. In summary, this study harnesses learner corpus techniques and computational tools to elucidate empirical patterns in the emergence of cohesive competence in novice writing.

IV. Results and Discussion

According to Crossley et el. (2016), TAACO calculates a variety of local, global, and overall text cohesion markers classified into five categories, namely type–token ratio (TTR), lexical overlap, semantic overlap, connectives and Givenness which have been proved to demonstrate positive relations with measures of cohesion in previous research (McNamara et al., 2010; Crossley & McNamara, 2011). The program has been validated on a corpus of writing by L2 college students. In this study, we are trying to calculate the values for the five groups of indices for EFL beginner writing to determine the effective indices for significant relations that mark progress and development.

A. Type–Token Ratio (TTR)

The type-token ratio (TTR) is a measure that quantifies the repetition of words in a text, which is calculated by dividing the number of unique words (types) by the total number of words (tokens). The TAACO program calculates several TTR indices, including simple TTR (the ratio of types to tokens for all words) and content word TTR (the ratio using only nouns, verbs, adjectives, and adverbs). In addition, TAACO computes lemma TTR, which uses lemmas instead of word forms, and content lemma TTR. Beyond traditional word-based TTR, TAACO also determines TTR for bigrams (two-word strings) and trigrams (three-word strings). These lexical diversity metrics capture repetition at the word and phrase level.

TAACO calculates altogether 15 indices for TTR. Lemma TTR (lemma-ttr) is measured by the number of unique lemmas (types) divided by the number of total running lemmas (tokens). Lemma MATTR (lemma-mattr) is the moving average TTR with 50-word window. Lexical density of tokens (lexical-density-tokens) is the percentage of text tokens that are content words. Lexical density of types (lexical-density-types) is the percentage of text types that are content words. Content lemma TTR (content-ttr) is the number of unique content word lemmas (types) divided by the number of total content word lemmas (tokens). Function lemma TTR (function-ttr) is measured by the number of unique function word lemmas (types) divided by the number of total function word lemmas (tokens). Function word MATTR (function-mattr) is the moving average function word TTR with 50-word window. Noun lemma TTR (noun-ttr) is measured by the number of unique noun lemmas (types) divided by the number of total noun lemmas (tokens). Verb lemma TTR (verb-ttr) is measured by the number of unique verb lemmas (types) divided by the number of total verb lemmas (tokens). Adjective lemma TTR (adj-ttr) is measured by the number of unique adjective lemmas (types) divided by the number of total adjective lemmas (tokens). Adverb lemma TTR (adv-ttr) is measured by the number of unique adverb lemmas (types) divided by the number of total adverb lemmas (tokens). Pronoun lemma TTR (prp-ttr) is measured by the number of unique pronoun lemmas (types) divided by the number of total pronoun lemmas (tokens). Argument lemma TTR (argument-ttr) is measured by the number of unique noun and pronoun lemmas (types) divided by the number of total noun and pronoun lemmas (tokens). Bigram lemma TTR (bigram-lemma-ttr) is measured by the number of unique bigram lemmas (types) divided by the number of total bigram lemmas (tokens). Trigram lemma TTR (trigram-lemma-ttr) is measured by the number of unique trigram lemmas (types) divided by the number of total trigram lemmas (tokens).

Table 1 shows the descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for TTR over the years. For indices like lexical density of tokens and pronoun lemma TTR, there are significant differences existing among the three year-group pairs, which demonstrates their apparent effectiveness in marking the development of beginner English. For another eight indices: lemma TTR, function lemma TTR, lexical density of types, noun lemma TTR, lemma MATTR, function word MATTR, argument lemma TTR and bigram lemma TTR, there are significant differences among some of the year groups, but not all, which indicates their partial effectiveness in predicting development of beginner English writing. For the remaining five indices: content lemma TTR, verb lemma TTR, adjective lemma TTR, adverb lemma TTR and trigram lemma TTR, there is no significant difference among the year groups, which means they are really ineffective for describing the development of beginner English writing.

TABLE 1
DESCRIPTIVE STATISTICS, RESULTS OF KRUSKAL-WALLIS TESTS, AND PAIRWISE COMPARISONS FOR TTR (BY YEAR)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year –2nd Year (Sig.)	2nd Year – 3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
lemma-ttr	1st Year	124	.6042056	.06154838	.007*	.220*	0.180*	1.000
	2nd Year	165	.5809601	.06316105				
	3rd Year	199	.6058319	.09135382				
	Total	488	.5970091	.07632501				
lemma-mattr	1st Year	124	.7330383	.05151894	.000*	1.000	.000*	.000*
	2nd Year	165	.7387137	.04365904				
	3rd Year	199	.7724083	.05883193				
	Total	488	.7510118	.05513662				
lexical-density-tokens	1st Year	124	.5352502	.04606519	.000*	.400*	.000*	.000*
	2nd Year	165	.5487304	.04813057				
	3rd Year	199	.5780773	.06033019				
	Total	488	.5572724	.05586542				
lexical-density-types	1st Year	124	.6890586	.04532788	.000*	.000*	1.00	.000*
	2nd Year	165	.7191128	.04311203				
	3rd Year	199	.7211756	.05527885				
	Total	488	.7123172	.05072148				
content-ttr	1st Year	124	.7796973	.08488925	.072	-	-	-
	2nd Year	165	.7630607	.08333150				
	3rd Year	199	.7567638	.10737782				
	Total	488	.7647203	.09449971				
function-ttr	1st Year	124	.4312852	.06487324	.000*	.000*	.000*	. 560
	2nd Year	165	.3897842	.07198707				
	3rd Year	199	.4308936	.09949560				
	Total	488	.4170934	.08493900				
function-mattr	1st Year	124	.4596455	.06219964	.000*	.056	.000*	.003*
	2nd Year	165	.4442500	.05251929				
	3rd Year	199	.4932052	.07839894				
	Total	488	.4681252	.06980977				
noun-ttr	1st Year	124	.8101165	.11615705	.000*	.030*	.131	.000*
	2nd Year	165	.7809462	.09964661				
	3rd Year	199	.7574599	.12028006				
	Total	488	.7787809	.11432796				
verb-ttr	1st Year	124	.7813169	.10521046	.135	-	-	-
	2nd Year	165	.7547443	.11438665				
	3rd Year	199	.7512365	.14670986				
	Total	488	.7600659	.12692249				
adj-ttr	1st Year	124	.8581928	.11626879	.423	-	-	-
	2nd Year	165	.8737070	.12184008				
	3rd Year	199	.8582106	.14901874				
	Total	488	.8634456	.13227205				
adv-ttr	1st Year	124	.7251454	.19014802	.247	-	-	-
	2nd Year	165	.7512745	.18258782				
	3rd Year	199	.7540799	.20661678				
	Total	488	.7457792	.19458221				
prp-ttr	1st Year	124	.2938143	.07626773	.000*	.000*	.000*	.002*
	2nd Year	165	.2367947	.09867630				
	3rd Year	199	.3966275	.20452447				
	Total	488	.3164610	.16328528				
argument-ttr	1st Year	124	.5969056	.09103748	.000*	1.000	.000*	.000*
	2nd Year	165	.5991390	.08992661				
	3rd Year	199	.6706779	.12953046				
	Total	488	.6277441	.11362483				
bigram-lemma-ttr	1st Year	124	.9266178	.04087659	.000*	.088	.000*	.393
	2nd Year	165	.9186062	.03589195				
	3rd Year	199	.9319951	.04642455				
	Total	488	.9261018	.04201648				
trigram-lemma-ttr	1st Year	124	.9856652	.01585286	.345	-	-	-
	2nd Year	165	.9839646	.01628005				
	3rd Year	199	.9847113	.02002056				
	Total	488	.9847012	.01777949				

The statistical significance detected among the ten lexical indices furnishes quantitative evidence that novice learners actively absorb new linguistic symbols and expand their vocabulary over time. This underscores the critical priority for beginners to devote focused study to continuously enlarging their vocabulary, particularly regarding function words and collocations, as this represents a foundational task undergirding communicative ability. For any language learner, the preliminary and most essential step is comprehending and utilizing the words that denote surrounding people, objects and ideas, since this lexical knowledge delineates the scope and sophistication of verbal and written expression. Hence, pedagogical materials and activities for novice English learners should be purposefully designed to target this core need for vocabulary enrichment. Simply put, constructing a basic vocabulary foundation should be considered the cornerstone in introductory courses for beginners across linguistic and situational contexts.

The statistically significant differences observed in pronoun-related indices, such as pronoun lemma type-token ratio, offer compelling quantitative evidence for the developmental trajectory of cohesive devices in novice English learners' writing. As demonstrated extensively in prior discourse analytic research, pronouns serve a vital cohesive role by establishing logical connections across sentences and paragraphs and avoiding repetitive use of proper nouns (Halliday & Hasan, 1976; Biber et al., 1999; Flowerdew, 2000). The skilled use of pronouns to maintain coherence is gradually and dynamically developed over time as learners gain proficiency. Similar to the acquisition process of other linguistic features, learners' mastery of pronominal cohesion is shaped by multiple factors, most saliently the semantic transparency and perceptual salience of specific pronoun forms and functions (Ellis & Simpson-Vlach, 2009; Goldberg, 2019).

Concordance analysis of the present learner corpus reveals a possible learning sequence progressing from personal and demonstrative pronouns toward increased use of possessive and reflexive pronouns. As personal and demonstrative pronouns tend to be more semantically and structurally transparent, as well as more grammatically salient, this aligns with usage-based theories which posit that transparency and salience facilitate acquisition (Tomasello, 2005; Ellis, 2002). The later emergence of reflexives and possessives accords with research showing that linguistically opaque features require more holistic processing and take longer to acquire (Glucksberg, 2001; Wray, 2000). Overall, the data-driven findings presented here support the conclusion that mastery of pronoun usage, and function words more broadly, serves as a critical step toward enhanced textual cohesion and readability in student writing. Moving forward, explicit instruction and targeted practice with opaque pronoun forms could speed the development of native-like cohesive device usage.

In contrast to the pronoun-related indices, the statistical insignificance of the lemma type-token ratios for content words, verbs, adjectives, adverbs, and trigrams provides insight into the persisting difficulties novice writers face. As demonstrated in prior discourse studies, content words like nouns, verbs, and adjectives constitute the primary means of expressing ideas, concepts, and emotions in a text (Biber & Conrad, 2019). The lack of sophisticated content word knowledge evidenced in learner writing may stem from underdeveloped vocabulary failing to provide the precise, expressive words required for descriptive, narrative, expository, and argumentative purposes (Nation, 2013). This deficiency is compounded by insufficient mastery of simple modifying structures, including adjectives and adverbs, which expand and refine lexical meanings (Schmid, 2012). Additionally, limited use of longer collocational bundles like trigrams may undermine the logical flow and connectivity of more complex ideas (Durrant & Schmitt, 2010).

The statistically flat growth trajectories for multiple type-token ratios illuminate ongoing content-word, modifier, and collocation usage gaps hindering learners' lexico-grammatical smoothness in writing. As evinced by the present corpus analysis, achieving native-like lexical richness, specificity, and collocational fluency remains an elusive benchmark for novice academic writers. Explicit vocabulary expansion interventions may help expedite mastery of the sophisticated content words and collocations essential to precise expression in writing (Crossley & Salsbury, 2010). Additionally, awareness-raising around modifier usage could sharpen learners' ability to write expressively and vividly.

B. Lexical Overlap (Sentence)

Lexical overlap, the repetition of words or phrases across different text parts, establishes cohesion and unity within writing. When specific terminology or expressions are echoed throughout a work, this signals to readers that these concepts are fundamentally connected and vital to the central theme. The strategic repetition of critical terms facilitates comprehension by reinforcing principal ideas, emphasizing salient points, and delineating a logical progression of thoughts. Moreover, lexical overlap aids in smoothly transitioning between sections or paragraphs, promoting fluidity by linking related concepts using repetitious language. Consequently, this technique enhances readability and strengthens the writer's material handling.

While lexical repetition may manifest within sentences, between sentences, or across paragraphs, novice writers often compose texts with minimal overlap due to their reliance on short or single paragraphs. Therefore, calculating lexical overlap indices across paragraphs proves insignificant mainly. This research consequently focuses exclusively on lexical overlap across sentences, which serves as a more meaningful measure of cohesion for fledgling beginner writing. This research utilizes TAACO to calculate six lexical overlap indices across sentences systematically: 1) adjacent sentence overlap - the average number of repeated words between consecutive sentences; 2) binary adjacent sentence overlap - the proportion of adjacent sentences containing any overlapping words; 3) adjacent two-sentence overlap - the average repeated words between a sentence and the sentence two units later; 4) adjacent sentence overlap (sentence normed) - the average repeated words between sentences divided by the number of sentences; 5) adjacent two-sentence overlap (sentence normed) - the average repeated words between a sentence divided by the number of sentences two units later; and 6) binary two-sentence overlap - the proportion of sentences containing overlapping words with the sentences two units later. These precise, automated measurements provide granular insights into patterns of lexical repetition between sentences in novice writing samples. The indices enable us to pinpoint strengths and deficiencies in students' utilization of lexical overlap to promote cohesion.

Table 2 shows the descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for the first part of the exact indices considered: adjacent sentence overlap all lemmas (adjacent overlap all sent), adjacent two-sentence overlap all lemmas (adjacent overlap 2 all sent), adjacent two-sentence overlap all lemmas (sentence normed) (adjacent overlap 2 all sent div seg) and binary adjacent two-sentence overlap all lemmas (adjacent overlap binary 2 all sent). The results of the Kruskal-Wallis tests indicate significant differences among the three year pairs for all six lexical overlap indices. This suggests that these measures can potentially track progression in students' utilization of repetition for cohesion in early writing. However, subsequent pairwise comparisons reveal that no indices differ significantly across all year pairs, implying these metrics alone may not fully capture developmental trajectories in novices' grasp of cohesion.

Table 2
Descriptive Statistics, Results of Kruskal-Wallis Tests, and Pairwise Comparisons for Lexical Overlap (Sentence, by Year, Part 1)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year –2nd Year (Sig.)	2nd Year – 3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
adjacent overlap all sent	1st Year	124	.1578766	.05075178	.003*	1.000	.003*	.081
	2nd Year	165	.1592696	.04736822				
	3rd Year	199	.1425744	.05764924				
	Total	488	.1521075	.05310691				
adjacent overlap all sent div seg	1st Year	124	1.4707991	.68918098	.006*	.043*	1.000	.005*
	2nd Year	165	1.7282382	.89807024				
	3rd Year	199	1.8897127	1.16017040				
	Total	488	1.7286705	.98302441				
adjacent overlap binary all sent	1st Year	124	.6742909	.16993930	.004*	.003*	.156	.302
	2nd Year	165	.7414923	.16452732				
	3rd Year	199	.6993865	.21838704				
	Total	488	.7072463	.19104805				
adjacent overlap 2 all sent	1st Year	124	.2454301	.06253737	.020*	1.000	.072	.046*
	2nd Year	165	.2406589	.05670242				
	3rd Year	199	.2268673	.07620190				
	Total	488	.2362472	.06704743				
adjacent overlap 2 all sent div seg	1st Year	124	2.3033885	1.09968501	.000*	.167	.104	.000*
	2nd Year	165	2.5690405	1.19281005				
	3rd Year	199	2.9563375	1.64827036				
	Total	488	2.6594734	1.40007217				
adjacent overlap binary 2 all sent	1st Year	124	.8135530	.13537375	.000*	.000*	1.000	.004*
	2nd Year	165	.8737486	.12998383				
	3rd Year	199	.8462186	.17929961				
	Total	488	.8472266	.15464910				

Several factors may explain the dynamic patterns in lexical repetition revealed by the indices. First, fledgling writers often rely on simple syntactic structures and short sentences, which intrinsically limits their capability to repeat words across sentences. Second, the indices quantify overlap across adjacent sentences, whereas students may begin repeating words in a more long-range, sophisticated manner before mastering repetition across consecutive sentences. Finally, the skill development in using lexical repetition for cohesion probably progresses non-linearly, with regressions as students experiment with repetition.

Table 3 shows the descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for the second set of the exact indices considered: adjacent sentence overlap content lemmas (adjacent overlap cw sent), adjacent sentence overlap content lemmas (sentence normed) (adjacent overlap cw sent div seg), binary adjacent sentence overlap content lemmas (adjacent overlap binary cw sent), adjacent two-sentence overlap content lemmas (adjacent overlap 2 cw sent), adjacent two-sentence overlap content lemmas (sentence normed) (adjacent overlap 2 cw sent div seg), binary adjacent two-sentence overlap content lemmas (adjacent overlap binary 2 cw sent), adjacent sentence overlap function lemmas (adjacent overlap fw sent), adjacent sentence overlap function lemmas (sentence normed) (adjacent overlap fw sent div seg), binary adjacent sentence overlap function lemmas (adjacent overlap binary fw sent), adjacent two-sentence overlap function lemmas (adjacent overlap 2 fw sent), adjacent two-sentence overlap function lemmas (sentence normed) (adjacent overlap 2 fw sent div seg), binary adjacent two-sentence overlap function lemmas (adjacent overlap binary 2 fw sent). Analysis of the lexical overlap indices for content and function words reveals nuanced insights into novice writers' evolving utilization of repetition for cohesive purposes. Regarding content word repetition, the indices show ambiguous developmental patterns, with only two of the six metrics indicating significant differences between one or two year pairs, but no indices differing significantly for all group pairs. This suggests fledgling writers' repetition of content words progresses irregularly rather than linearly. For function words, all six indices significantly differ between pairs, yet post hoc tests reveal significance for just one or two year pairs per index, not across all year pairs.

TABLE 3
DESCRIPTIVE STATISTICS, RESULTS OF KRUSKAL-WALLIS TESTS, AND PAIRWISE COMPARISONS FOR LEXICAL OVERLAP (SENTENCE, BY YEAR, PART 2)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year –2nd Year (Sig.)	2nd Year – 3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
adjacent overlap cw sent	1st Year	124	.0780125	.05507522	.743	-	-	-
	2nd Year	165	.0716493	.04742333				
	3rd Year	199	.0770217	.05802963				
	Total	488	.0754570	.05385292				
adjacent overlap cw sent div seg	1st Year	124	.4059277	.32269355	.002*	.683	.070	.003*
	2nd Year	165	.4743773	.39554561				
	3rd Year	199	.6278599	.58403441				
	Total	488	.5195725	.47585657				
adjacent overlap binary cw sent	1st Year	124	.3020822	.20292772	.013*	.898	.169	.014*
	2nd Year	165	.3270068	.20142620				
	3rd Year	199	.3825004	.25101926				
	Total	488	.3433031	.22546036				
adjacent overlap 2 cw sent	1st Year	124	.1242944	.07056928	.285	-	-	-
	2nd Year	165	.1163540	.06667118				
	3rd Year	199	.1275195	.07909610				
	Total	488	.1229248	.07295205				
adjacent overlap 2 cw sent div seg	1st Year	124	.6551183	.52264172	.000*	.424	.002*	.000*
	2nd Year	165	.7558467	.56955860				
	3rd Year	199	1.0204739	.81959010				
	Total	488	.8381633	.68974319				
adjacent overlap binary 2 cw sent	1st Year	124	.4274535	.21460352	.000*	.877	.003*	.000*
	2nd Year	165	.4656173	.25434359				
	3rd Year	199	.5438433	.27090414				
	Total	488	.4878195	.25618756				
adjacent overlap fw sent	1st Year	124	.2352727	.07576343	.000*	.006*	.000*	.882
	2nd Year	165	.2634762	.07884356				
	3rd Year	199	.2241963	.08710433				
	Total	488	.2402919	.08320001				
adjacent overlap fw sent div seg	1st Year	124	.9951241	.48049352	.016*	.013*	.803	.156
	2nd Year	165	1.2025078	.64480530				
	3rd Year	199	1.1923788	.77568986				
	Total	488	1.1456815	.67135207				
adjacent overlap binary fw sent	1st Year	124	.5881450	.17951129	.000*	.000*	.000*	1.000
	2nd Year	165	.6753525	.18334110				
	3rd Year	199	.5952122	.23114444				
	Total	488	.6205131	.20670729				
adjacent overlap 2 fw sent	1st Year	124	.3596506	.08641044	.000*	.008*	.000*	1.000
	2nd Year	165	.3911772	.08740202				
	3rd Year	199	.3524277	.11276817				
	Total	488	.3673648	.09963449				
adjacent overlap 2 fw sent div seg	1st Year	124	1.5369767	.69985564	.029*	.084	1.000	.037*
	2nd Year	165	1.7418699	.80131814				
	3rd Year	199	1.8426520	1.06607109				
	Total	488	1.7309045	.90373177				
adjacent overlap binary 2 fw sent	1st Year	124	.7478181	.15854199	.001*	.001*	.045	.451
	2nd Year	165	.8146724	.14788899				
	3rd Year	199	.7547786	.21638209				
	Total	488	.7732609	.18339906				

The same factors may explain these complex, nonlinear patterns. On the one hand, novice writers may rely heavily on short sentences and simple structures, distorting index scores. The descriptive statistics for novice writing indicate scores of 3.8 for the number of letters per word and 14.0 for the number of words per sentence, while that for the writing of native speakers (LOCNESS) goes up to 4.7 for the number of letters per word and 20.2 for the number of words per sentence. On the other, contextual factors like assignment type likely influence repetitive patterns. College beginners in their freshman year are mainly required to complete writing assignments of letters, notes or emails; while in the sophomore year, expositive or narrative essays are required; and in the third argumentations form the major tasks.

C. Semantic Overlap

TAACO 2.0 calculates the average similarity between progressive adjacent segments (sentences or paragraphs) in a text. But because novice writers may usually tend to have only one paragraph for a piece of writing, semantic similarity is calculated at the sentence level in this study. Table 4 shows the descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for semantic overlap among sentences, and the exact indices considered include: average sentence-to-sentence overlap of noun synonyms (syn overlap sent noun), average sentence-to-sentence overlap of verb synonyms (syn overlap sent verb), average latent semantic analysis cosine similarity between all adjacent sentences (lsa 1 all sent), average latent semantic analysis cosine similarity between all adjacent sentences (with a two-sentence span) (lsa 2 all sent), average latent dirichlet allocation divergence score between all adjacent sentences (lda 1 all sent), average latent dirichlet allocation divergence score between all adjacent sentences (with a two-sentence span) (lda 2 all sent), average word2vec similarity score between all adjacent sentences (word2vec 1 all sent) and average word2vec similarity score between all adjacent sentences (with a two-sentence span) (word2vec 2 all sent).

TABLE 4
DESCRIPTIVE STATISTICS, RESULTS OF KRUSKAL-WALLIS TESTS, AND PAIRWISE COMPARISONS FOR SEMANTIC OVERLAP (BY YEAR)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year –2nd Year (Sig.)	2nd Year – 3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
syn overlap sent noun	1st Year	124	.2348546	.27146316	.000*	.320	.000*	.000*
	2nd Year	165	.2923775	.30947867				
	3rd Year	199	.5030662	.56995730				
	Total	488	.3636771	.44365290				
syn overlap sent verb	1st Year	124	.1905886	.19879583	.047*	.854	.454	.045*
	2nd Year	165	.2460485	.28728015				
	3rd Year	199	.2904373	.31939810				
	Total	488	.2500574	.28429275				
lsa 1 all sent	1st Year	124	.2641947	.11975273	.186	-	-	-
	2nd Year	165	.2440322	.09666080				
	3rd Year	199	.2434589	.12981950				
	Total	488	.2489217	.11704729				
lsa 2 all sent	1st Year	124	.6180059	.08201105	.842	-	-	-
	2nd Year	165	.6164877	.09359811				
	3rd Year	199	.5968245	.12774386				
	Total	488	.6088550	.10669454				
lda 1 all sent	1st Year	124	.9336287	.04852075	.001*	.183	.148	.001*
	2nd Year	165	.9454044	.04305631				
	3rd Year	199	.9493092	.05285490				
	Total	488	.9440046	.04893147				
lda 2 all sent	1st Year	124	.9626876	.03757985	.064	-	-	-
	2nd Year	165	.9633775	.08165994				
	3rd Year	199	.9635150	.10243107				
	Total	488	.9632583	.08287458				
word2vec 1 all sent	1st Year	124	.7651502	.04536008	.000*	.000*	.326	.000*
	2nd Year	165	.7892150	.04556824				
	3rd Year	199	.7876728	.09221712				
	Total	488	.7824713	.06913218				
word2vec 2 all sent	1st Year	124	.8213032	.03238955	.000*	.001*	1.000	.000*
	2nd Year	165	.8329795	.05591308				
	3rd Year	199	.8181638	.11719341				
	Total	488	.8239709	.08333550				

The results of the Kruskal-Wallis tests indicate that for some of the indices for semantic overlap there are significant differences (five out of eight). Post hoc pairwise comparisons revealed that statistically significant differences exist for one or two of the year pairs but none for all three pairs. This denotes that indices for semantic overlap help predict cohesion development of novice writers, but they are not the most reliable. The results also confirm the dynamism of cohesion development of novice writing. It may be explained by the fact that novice writers struggle with the right words to express themselves. At the beginning of their English learning journey, they are trying to pick the everyday linguistic forms for things around them, while complex semantic relations like synonymy, polysemy, antonymy and hyponymy may only develop later with their improvement in vocabulary richness. Further concordance analysis of high-frequency simple verbs like “think” and their synonyms like “state”, “claim”, “argue”, “hold”, “believe”, "announce" and structures with a similar sense of expressing opinions like "in … opinion", "form … perspective", "take…for granted" and "point out", the search finds that the most frequently used is "think" (Fre. = 103), followed by believe (20) and argue (2). Most strikingly, other expressions and structures on the above list find no occurrence in the corpus.

D. Connectives

Connectives are the standard means for a writer to ensure cohesion and coherence in writing. TAACO 2.0.4 calculates 25 connectives indices, for each of which the occurrence of each item is counted and the sum is divided by the total number of words in the text. The indices for this study are mainly considered on the sentence level. Table 5 shows the descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for the first half of indices, including basic connectives, conjunctions, disjunctions, lexical subordinators, coordinating conjuncts, addition, sentence linking, order, reason and purpose, all causal connectives, positive causal connectives and opposition. For basic connectives, it calculates the frequency of words like "for", "and", "nor"; for conjunctions, words like "and", "but"; for disjunctions, words like "or"; for lexical subordinators, words like "after", "although", "as"; for coordinating conjuncts, words like "yet", "so", "nor"; for addition, words like "and", "also", "besides"; for sentence linking, words like "nonetheless", "therefore", "although"; for order, words like "to begin with", "next", "first"; for reason and purpose, words like "therefore", "that is why", "for this reason"; for all causal connectives, words like "although", "arise", "arises". From Table 5, the results of the Kruskal-Wallis tests indicate that for some of the indices for the first set of connectives there are significant differences (nine out of twelve), and post hoc pairwise comparisons reveal that the statistically significant differences exist for one or two of the year pairs, but none for all three pairs, which denotes that indices for connectives help to predict cohesion development of novice writes, but they are not the most reliable. The results also confirm the dynamism of cohesion development of novice writing. It may be explained by the fact that connectives are the standard practice for cohesion despite languages. It's mainly a positive transfer for novice writers to transfer the connections from their mother tongue to a second language they are using.

TABLE 5
DESCRIPTIVE STATISTICS, RESULTS OF KRUSKAL-WALLIS TESTS, AND PAIRWISE COMPARISONS FOR CONNECTIVES (BY YEAR, PART 1)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year –2nd Year (Sig.)	2nd Year –3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
basic connectives	1st Year	124	.0387581	.01814382	.471	-	-	-
	2nd Year	165	.0410173	.01786714				
	3rd Year	199	.0411798	.02088882				
	Total	488	.0405095	.01921210				
conjunctions	1st Year	124	.0266556	.01407599	.037*	.121	1.000	.042*
	2nd Year	165	.0303149	.01481850				
	3rd Year	199	.0313912	.01818129				
	Total	488	.0298240	.01618723				
disjunctions	1st Year	124	.0022405	.00452989	.001*	1.000	.008*	.005*
	2nd Year	165	.0023572	.00469849				
	3rd Year	199	.0039463	.00566858				
	Total	488	.0029755	.00513160				
lexical subordinators	1st Year	124	.0123730	.01141277	.780	-	-	-
	2nd Year	165	.0119932	.01028290				
	3rd Year	199	.0110394	.00874205				
	Total	488	.0117007	.00999497				
coordinating conjuncts	1st Year	124	.0100875	.01053715	.013*	1.000	.109	.018*
	2nd Year	165	.0087769	.00877796				
	3rd Year	199	.0068392	.00762666				
	Total	488	.0083197	.00891075				
addition	1st Year	124	.0266154	.01505393	.046*	.056	.708	.040*
	2nd Year	165	.0292851	.01513882				
	3rd Year	199	.0316869	.01918912				
	Total	488	.0295861	.01697662				
sentence linking	1st Year	124	.0172107	.01383554	.010*	.033*	1.000	.014*
	2nd Year	165	.0203959	.01115989				
	3rd Year	199	.0210441	.01275674				
	Total	488	.0198509	.01260586				
order	1st Year	124	.0049391	.00701321	.076	-	-	-
	2nd Year	165	.0031417	.00515463				
	3rd Year	199	.0037354	.00558599				
	Total	488	.0038405	.00587655				
reason and purpose	1st Year	124	.0067048	.00914589	.001*	.001*	.664	.027*
	2nd Year	165	.0101053	.00964415				
	3rd Year	199	.0086312	.00821996				
	Total	488	.0086401	.00903328				
all causal	1st Year	124	.0138304	.01277783	.009*	.038*	.016*	1.000
	2nd Year	165	.0170166	.01232715				
	3rd Year	199	.0133305	.01059667				
	Total	488	.0147038	.01186765				
positive causal	1st Year	124	.0136920	.01190992	.000*	.000*	.013*	.021*
	2nd Year	165	.0212769	.01321436				
	3rd Year	199	.0172298	.01191549				
	Total	488	.0176993	.01268343				
opposition	1st Year	124	.0045854	.00686894	.002*	.161	.277	.001*
	2nd Year	165	.0056429	.00609159				
	3rd Year	199	.0069109	.00648411				
	Total	488	.0058913	.00651050				

Table 6 shows the descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for the second half of indices for connectives, including determiners (a, an, the), demonstratives (this, that, these), attended demonstratives (this + noun), unattended demonstratives (this as the subject), all additive connectives (after all, again, all in all), all logical connectives (actually, admittedly, after all), positive logical connectives (actually, after all, all in all), negative logical connectives (admittedly, alternatively, although), temporal connectives (a consequence of, after, again), positive intentional connectives (by, desire, desired), all positive connectives (actually, after, again), all negative connectives (admittedly, alternatively, although), all connectives (actually, admittedly, after). The results of the Kruskal-Wallis tests indicate that for some of the indices for the second set of connectives there are significant differences (nine out of thirteen), and post-hoc pairwise comparisons reveal that the statistically significant differences exist for one or two of the three year pairs, but none for all three pairs, which denotes that indices for connectives help to predict cohesion development of novice writing but not the most reliable, which also confirms the dynamism of cohesion development of novice writing.

TABLE 6
DESCRIPTIVE STATISTICS, RESULTS OF KRUSKAL-WALLIS TESTS, AND PAIRWISE COMPARISONS FOR CONNECTIVES (BY YEAR, PART 2)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year –2nd Year (Sig.)	2nd Year –3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
determiners	1st Year	124	.0561664	.02296447	.000*	.019*	.004*	.000
	2nd Year	165	.0647777	.02538318				
	3rd Year	199	.0732913	.02862552				
	Total	488	.0660613	.02702165				
all demonstratives	1st Year	124	.0143748	.00994088	.079	-	-	-
	2nd Year	165	.0130695	.01093394				
	3rd Year	199	.0123391	.01221077				
	Total	488	.0131033	.01124678				
attended demonstratives	1st Year	124	.0061021	.00746829	.050	-	-	-
	2nd Year	165	.0070970	.00935783				
	3rd Year	199	.0044009	.00583952				
	Total	488	.0057447	.00767010				
unattended demonstratives	1st Year	124	.0082726	.00719268	.015*	.011*	.513	.236
	2nd Year	165	.0059725	.00680695				
	3rd Year	199	.0084828	.01532693				
	Total	488	.0075806	.01120313				
all additive	1st Year	124	.0357418	.01847137	.000*	.283	.065	.000*
	2nd Year	165	.0394300	.01850796				
	3rd Year	199	.0443954	.02098351				
	Total	488	.0405176	.01982162				
all logical	1st Year	124	.0249926	.01625636	.001*	.029*	.808	.001*
	2nd Year	165	.0296871	.01373013				
	3rd Year	199	.0318111	.01702551				
	Total	488	.0293604	.01598216				
positive logical	1st Year	124	.0127967	.01220073	.415	-	-	-
	2nd Year	165	.0132311	.01032768				
	3rd Year	199	.0133554	.00932802				
	Total	488	.0131714	.01043755				
negative logical	1st Year	124	.0042929	.00610710	.002*	.071	.571	.001*
	2nd Year	165	.0056675	.00593446				
	3rd Year	199	.0066910	.00627052				
	Total	488	.0057356	.00617821				
all temporal	1st Year	124	.0106434	.01067377	.001*	.001*	.713	.015*
	2nd Year	165	.0059803	.00635626				
	3rd Year	199	.0072160	.00763829				
	Total	488	.0076691	.00833090				
positive intentional	1st Year	124	.0061354	.00779197	.001*	.009*	.003*	1.000
	2nd Year	165	.0093846	.00977897				
	3rd Year	199	.0060902	.00714807				
	Total	488	.0072156	.00840953				
all positive	1st Year	124	.0567093	.02402958	.304	-	-	-
	2nd Year	165	.0610915	.02334297				
	3rd Year	199	.0592777	.02386800				
	Total	488	.0592383	.02374325				
all negative	1st Year	124	.0066466	.00767763	.000*	.367	.016*	.000*
	2nd Year	165	.0080626	.00777032				
	3rd Year	199	.0106373	.00848917				
	Total	488	.0087527	.00820137				
all connective	1st Year	124	.0477623	.02099567	.000*	.016*	.214	.000*
	2nd Year	165	.0557663	.02290641				
	3rd Year	199	.0603775	.02406454				
	Total	488	.0556129	.02341914				

E. Givenness

Givenness indices approximate the proportion of given information to new information by examining pronoun density, pronoun-to-noun ratios, and repeated content lemmas and pronouns. TAACO 2.0.4 calculates four indices related to givenness: pronoun density, pronoun-to-noun ratio, repeated content lemmas, repeated content lemmas and pronouns. Table 7 shows the descriptive statistics, results of Kruskal-Wallis tests, and pairwise comparisons for the indices for givenness.

TABLE 7
DESCRIPTIVE STATISTICS, RESULTS OF KRUSKAL-WALLIS TESTS, AND PAIRWISE COMPARISONS FOR GIVENNESS (BY YEAR)

Variables	Year	N	Mean	Std. Deviation	Kruskal-Wallis test (Sig.)	Pairwise Comparisons		
						1st Year –2nd Year (Sig.)	2nd Year – 3rd Year (Sig.)	1st Year – 3rd Year (Sig.)
pronoun density	1st Year	124	.0245356	.01648155	.000*	.343	.000*	.000*
	2nd Year	165	.0214381	.01753344				
	3rd Year	199	.0343214	.02258267				
	Total	488	.0274788	.02032674				
pronoun noun ratio	1st Year	124	.1026146	.07459447	.000*	.015*	.000*	.390
	2nd Year	165	.0803862	.07104280				
	3rd Year	199	.1192520	.09917957				
	Total	488	.1018834	.08593164				
repeated content lemmas	1st Year	124	.1930637	.06792846	.026*	.417	.623	.021*
	2nd Year	165	.2063716	.06946115				
	3rd Year	199	.2151526	.08818968				
	Total	488	.2065709	.07764562				
repeated content and pronoun lemmas	1st Year	124	.2075686	.06647654	.000*	.345	.027*	.000*
	2nd Year	165	.2215994	.07464513				
	3rd Year	199	.2413953	.08882318				
	Total	488	.2261067	.07993782				

The results of the Kruskal-Wallis tests indicate that for all the indices, there are significant differences, and post hoc pairwise comparisons reveal that the statistically significant differences exist for one or two of the year pairs, but none for all three pairs, which denotes that indices for givenness help to predict cohesion development of novice writing, but they are not the most reliable. The results also confirm the dynamism of cohesion development of beginners. The explanation could be the repetition of pronouns in beginners' writing is commonly and frequently utilized. Therefore, the high repetition and frequency of pronouns contribute to the high score of pronoun-related indices.

V. CONCLUSION

This corpus-based study explored the development of cohesive device usage in beginner EFL writing over time through quantitative analysis of a longitudinal learner corpus. The results reveal noteworthy insights into novice writers' evolving mastery of cohesion and discourse competence. Most strikingly, the statistical analyses demonstrate that indices related to pronouns, including pronoun density, pronoun-noun ratios, and pronoun repetition, differ significantly across year groups and reliably track progression in cohesive proficiency. This aligns with previous research emphasizing pronouns' vital cohesive role in establishing logical connections and maintaining coherence (Flowerdew, 2000). It also provides quantitative confirmation of the critical importance of function words in developing native-like written discourse (Biber et al., 1999; Goldberg, 2019). Based on the corpus data, learners appear to acquire personal and demonstrative pronouns earlier, followed by possessive and reflexive forms, consistent with usage-based acquisition patterns where transparent features emerge first (Goldberg, 2019; Ellis, 2002).

However, the study findings indicate that most lexical, syntactic and connective indices, though demonstrating some pairwise differences, do not reliably distinguish or predict writing development for beginners. The indices for content word diversity and overlap, conjunctions, and modifiers reveal ambiguous developmental trajectories over time. This contrasts prior research using similar NLP tools which found lexical bundles, cohesive markers and sophistication indices effectively predict essay scores, even for novice writers (Crossley et al., 2016; McNamara et al., 2010).

The divergence suggests persistent difficulties in mastering sophisticated content words, precise nuances, and appropriate collocations, which undermine beginners' lexical repetition and cohesive patterning (Nation, 2013; Durrant & Schmitt, 2010). Concordance analysis verifies learners' predominant reliance on high-frequency pronouns like "he" and "it", while more varied opinion-stating expressions remain scarce. The indices' instability over time indicates a complex, nonlinear acquisition process as novices experiment with forms.

In conclusion, this study's unique longitudinal perspective provides empirical insights into beginners' cohesion development, highlighting persistent hurdles like content-word gaps while tracking statistical shifts in pronoun usage. The findings reveal both universal sequencing principles and persistent challenges shaping the discourse competence of novice writers. They underscore the vital need for vocabulary expansion and explicit cohesion instruction targeting opaque features like possessive pronouns and precise lexico-grammatical expressions. The techniques used also demonstrate the value of learner corpora and computational tools in generating data-driven insights to inform instruction and assessment. Exploring lexical patterning development at the multi-word level could further enrich understanding of this critical dimension of EFL writing proficiency.

ACKNOWLEDGEMENTS

The present research is funded by the Innovative Practice Base for Developmental Integration of Information Technology with Foreign Languages Teaching and Research, Inner Mongolia University.

REFERENCES

[1] Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press.

[2] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education Limited.

[3] Bitchener, J., & Basturkmen, H. (2006). Perceptions of the difficulties of postgraduate L2 thesis students writing the discussion section. *Journal of English for Academic Purposes, 5*(1), 4-18. https://doi.org/10.1016/j.jeap.2005.10.002

[4] Connor, U. (1984). A study of cohesion and coherence in English as a second language students' writing. *Research in the Teaching of English*, *18*(3), 301-316. https://doi.org/10.1080/08351818409389208

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

[6] 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

[7] Crossley, S.A., & McNamara, D.S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236-1241).

[8] Crossley, S.A., & McNamara, D.S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66-79. https://doi.org/10.1016/j.jslw.2014.09.006

[9] Crossley, S.A., & Salsbury, T. (2010). The development of lexical bundle accuracy and production in English second language speakers. *IRAL - International Review of Applied Linguistics in Language Teaching*, *48*(1), 1-26. https://doi.org/10.1515/iral.2011.001

[10] Crossley, S.A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *51*, 1030-1046. https://doi.org/10.3758/s13428-017-0924-4

[11] Crossley, S.A., Roscoe, R., & McNamara, D.S. (2014). What is successful writing? An analysis of linguistic features of independent and integrated essay prompts. *Written Communication*, *31*(2), 202-221. https://doi.org/10.1177/0741088314526354

[12] Durrant, P., & Schmitt, N. (2010). Adult learners’ retention of collocations from exposure. *Second Language Research*, *26*(2), 163-188. https://doi.org/10.1177/0267658309349431

[13] Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143-188. https://doi.org/10.1017/S0272263102002024

[14] Ellis, R., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5*(1), 61-78. https://doi.org/10.1515/CLLT.2009.003

[15] Flowerdew, J. (2000). Signalling nouns in discourse. *English for Specific Purposes*, *22*(4), 329-346. https://doi.org/10.1016/S0889-4906(02)00017-0

[16] Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. Oxford University Press.

[17] Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

[18] Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). John Benjamins.

[19] Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, *15*(1), 17-27. https://doi.org/10.1111/j.1467-971X.1996.tb00089.x

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

[21] Hyland, K. (2004). *Genre and second language writing*. University of Michigan Press.

[22] Kong, K., & Pearson, P.D. (2003). The road to participation: The construction of a literacy practice in a learning community of linguistically diverse learners. *Research in the Teaching of English*, *38*(1), 85-124.

[23] McNamara, D.S., Crossley, S.A., & McCarthy, P.M. (2010). Linguistic features of writing quality. *Written Communication*, *27*, 57-86. https://doi.org/10.1177/0741088309351547

[24] McNamara, D.S., Crossley, S.A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. Behavior Research Methods, *45*(2), 499-515. https://doi.org/10.3758/s13428-012-0258-1

[25] Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 119-141). John Benjamins.

[26] Nation, P. (2013). *Learning vocabulary in another languag*e (2nd ed.). Cambridge University Press.

[27] Schmid, H. (2012). *English abstract nouns as conceptual shells: From corpus to cognition.* De Gruyter Mouton.

[28] Shin, Y. K., & Kim, Y. (2017). Using lexical bundles to teach articles to L2 English learners of different proficiencies. *System*, *69*, 79-91. https://doi.org/10.1016/j.system.2017.08.002

[29] Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

[30] Tsou, W. (2005). Improving speaking skills through instruction in oral classroom participation. *Foreign Language Annals*, *38*(1), 46-55. https://doi.org/10.1111/j.1944-9720.2005.tb02452.x

[31] Witte, S. P., & Faigley, L. (1981). Coherence, Cohesion, and Writing Quality. *College Composition and Communication*, *32*(2), 189–204. https://doi.org/10.2307/356693

[32] Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*(4), 463-489. https://doi.org/10.1093/applin/21.4.463

[33] Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, *20*(1), 1-28. https://doi.org/10.1016/S0271-5309(99)00015-4

[34] Zhang, M. (2000). Cohesive features in the expository writing of undergraduates in two Chinese universities. *RELC Journal*, *31*(1), 61-95. https://doi.org/10.1177/003368820003100104

Weilu Wang, born in Hohhot, Inner Mongolia, PRC. He is an associate professor in Foreign Languages College of Inner Mongolia University. His research interest includes corpus linguistic, applied linguistics and translation technology.

Wei Chen is working as a college lecturer at Inner Mongolia University. Born in April 1981, she was raised in a Daur family. Her research interests include applied linguistics and American literature.

Shuangshuang Shi was born in Ordos, China in 1980. She received her Master degree in English language and literature from Inner Mongolian University, China in 2010. She is currently a lecturer in the College of Foreign Languages, Inner Mongolian University. Her research interests include British literature and English teaching.

Lin Tong was born in Inner Mongolia, China in 1983. She received her Master's degree in English language and literature from Inner Mongolia University, China in 2010. She is currently a lecturer in the School of Foreign Languages, Inner Mongolia University, Hohhot, China. Her research interests include translation and American literature.

Manfu Duan, born in Hohhot, Inner Mongolia, PRC. He is a professor in Foreign Languages College of Inner Mongolia University and Chief of Division of International Cooperation and Exchange, Inner Mongolia University. His research interest includes translation study and applied linguistics.