Vol. 3 No. 4 (2025)

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

Syntactic Complexity and Lexical Density in EFL Academic Writing: A Corpus-Based Investigation of Developmental Patterns

Ayesha Saddique
Department of English, Govt. Post Graduate College for Women, Mandi Bahauddin, Punjab, Pakistan
Email: ayeshasaddique7778@gmail.com

Mohammad Aafaq Nadeem
MPhil Scholar, The University of Lahore Sargodha Campus
Email: iamaafaq84@gmail.com

Muhammad Danish
MPhil Scholar, The University of Lahore Sargodha Campus
Email: khawajadanish1769@gmail.com

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Abstract

This study investigates syntactic complexity and lexical density in academic writing by Arabic L-1 learners of English as Foreign Language (EFL) through a corpus-based methodology. A 350, 000-word dataset of argumentative essays was compiled from learners at CEFR B1, B2, and C1, evenly distributed across humanities and material sciences. Automated analysis was conducted using the L2 Syntactic Complexity Analyzer (L2SCA) and AntConc, with Stanford POS tagging applied to calculate 14 syntactic indices and lexical density via the Ure formula. Results revealed a progressive developmental trajectory: mean length of T-unit increased from 12.87 at B1 to 19.04 at C1, while complex nominals per clause rose by 84%, surpassing clause elaboration. Lexical density also advanced from 48.7 to 56.3 with a competitive relationship at B2 (r =.32) shifting to positive synergy at C1 (r =.56). Humanities learners at higher proficiency levels produced more nominal structures than science learners. Multiple regression identified complex nominals and lexical density as predictors of writing quality, accounting for 58% of variance. Findings show phrasal complexity as a key marker of the academic writing maturity and designate B2 as a critical stage of linguistic restructuring. The research contributes to second language development theory and informs EFL pedagogy by emphasizing nominalization, lexical sophistication and disciplinary writing performance.

Keywords: Syntactic Complexity, Lexical Density, EFL Academic Writing, Corpus-Based Analysis, Arabic L1, Writing Proficiency

Introduction

Academic writing in English as Foreign Language (EFL) settings represent one of the critical competencies of learners as they go through the process of higher education and work, though it is a cognitively and linguistically challenging area (Hyland, 2016). The primary issues in these problems are syntactic complexity, in terms of subordination of clauses, elaboration of phrases, and mean length of T-unit, and lexical density, measured as a ratio between content and total words that collectively predetermine the depth of information and structural complexity of academic speech (Biber et al., 2021). The hierarchical combination of ideas through syntactic complexity helps EFL authors build sophisticated arguments whereas the text conciseness through lexical density and the richness of the argument through semantic statuses are characteristic features of advanced academic texts (Lu, 2011). More recent studies with the help of corpus analyses have shown that EFL learners prefer clausal coordination over subordination in the early stages of learning because of the processing limitations and L1 transfer influences (Park, 2022). Such an uneven development of grammar highlights why syntactic and lexical development were interconnected, as early learners are more focused on lexical development, neglecting the need to deepen the format (O'Leary and Steinkrauss, 2022). In turn, it is obligatory to gain knowledge about these dimensions in order to identify the deficiencies in proficiency and then use specific instructional interventions in the EFL academic writing pedagogy.

Though significant progress has been made regarding the study of L2 writing, there remain important gaps in the systematic investigation of the phenomenon of syntactic complexity and lexical density in EFL-based corpus. Classical models like the multidimensional analysis offered by Biber (1988) have cast much lighter on register-specific linguistic patterns were academic writing favours nominalization and phrasal modification to clausal dependency. Nevertheless, these studies have been specifically focused on L1 English or on heterogeneous L2 samples, frequently excluding EFL learners who have a L1 Western (typically French, German, Spanish, and Italian) background (e.g., Sino-Tibetan and Arabic-speaking cohorts) (Shen et al., 2023). Furthermore, previous research is often based on small corpora (usually less than 50000 words) or cross-sectional designs, which is restrictive in terms of its ability to model development trajectories or discipline-specific differences (Wang et al., 2023). Even though the automated programs, such as the L2 Syntactic Complexity Analyzer (Lu, 2010), can be used to increase the accurateness of measurements, limited studies combine finer-grained indices of syntax (e.g., complex nominals per clause) and lexical density measures in EFL academic settings (Nasrabady et al., 2025). The given fragmentation gives a partial view of the development of these features in co-occurrence, notably when directed by cognitive load, where the competition process may incite the detrimental effect of syntactic elaboration to be lost in favor of lexical sophistication (O'Leary and Steinkrauss, 2022). The current research circumvents these shortcomings by applying the method on a scale of mass corpus that utilizes the representative EFL academic texts to provide strong and general information about the complexity of language.

Research Question[s]

Based on usage-based and developmental models of L2 language acquisition, the study presents a corpus-based approach to explore the relationship between the complexity of the syntactic structure and density of lexical knowledge in EFL academic compositional writing. The research is pegged into four research questions:

RQ1: What is the extent of variation in the syntactic complexity of argumentative EFL learning groups measured by mean length of T-unit (MLT), clauses per T-unit (C/T), and complex nominals per clause (CN/C)-across different CEFR proficiency levels in the argumentative essay writing?

RQ2: What is the relationship between lexical density across proficiency bands, and how these measures correlate with syntactic complexity?

RQ3: How does discipline setting (e.g. humanities vs. material sciences) moderate syntactic and lexical complexity in EFL academic writing, given that scientific registers are expected to contain higher nominal density consistent with register variation models?

RQ4: To what extent do aggregated complexity measures predict holistic ratings of EFL corpora writing quality, assuming that balanced linguistic profiles enhance coherence and persuasiveness?

The study has enormous implications on the process of acquiring a second language (SLA), corpus linguistics, and computer-assisted language learning (CALL). In SLA, the results improve developmental measures by outlining proficiency-based patterns of syntactic and lexical growth, especially in the context of the poorly represented EFL groups (Shen et al., 2023). Theoretically, the research is an extension of Processability Theory (Pienemann, 1998): corpus-based complexity measurements are combined, which means the interrelation between the paradigms of psycholinguistics

and computationalism. In CALL, intelligent tutoring systems can be based on automated reviews of learner corpus, which provides real-time feedback regarding the lack of complexity to promote autonomous writing progress (Park, 2022). From a pedagogical perspective, knowledge of the effects of trade-offs is put into scaffolded instructions protocols such as scaffold phrasal embedding to achieve higher lexical and sentence density but reduced syntactics (Ali, 2025; Wang et al., 2023). In the end, enhancing linguistic equity in scholarly communication worldwide by giving EFL writers the power to estimate native-like academic norms, this work paves the way to linguistic equity.

The paper proceeds as follows: the literature review is a synthesis of theoretical and empirical studies on the complexity of linguistics; the methodology summarizes the overview of the study in terms of corpus collection and analysis; the results present statistics and descriptive research findings, which are visualized by data table and graphs; the discussion compared the results with previous studies, limits are recognized, and the future is projected, and the conclusion provides an overview of the research and its practical application.

Theoretical Framework

Linguistic complexity in second language (L2) writing is based on structural and lexical aspects of textual production based on constructs. Syntactic complexity Syntactic complexity has been operationalized through syntactic developmental indices created by Hunt, which have been used to quantify the subordination, coordination, and phrasal elaboration of syntactic structures in terms of syntactic units (T-units). Mean length of T-unit (MLTU), clauses per T-unit (C/T), dependent clauses per clause (DC/C), and complex nominals per clause (CN/C) are the most important metrics that sum up to the hierarchical embedding that is inherent in academic discourse (Biber et al., 2021). In addition to this, lexical density, Ure (1971) expresses as lexical (content) words in the ratio of lexical word count to total word count times 100, measures both informational compactness and semantic load, and high scores reflect high propositional density. The theoretical basis of these constructions lies in Processability Theory (Pienemann, 1998), according to which grammatical development in L2 will occur in predictable stages, with processing prerequisites, so that the complexities of the phrasal level will be acquired after the clausal one. Usage-Based Linguistics (UBL) also sheds more light on this direction, arguing that entrenchment of multi-word units through frequency encourages nominal elaboration in expert-like language (Ellis, 2017). These systems coincide in EFL settings to the anticipation of a developmental change towards phrasal as opposed to clausal sophistication that is subject to cognitive resource and the exposure to input to create a potent corpus-focused analysis component.

Previous studies on Syntactic Complexity.

Inquiries based on corpora into the syntactic complexity have provided fine grained details in understanding developmental patterns of L2 and the seminal study by Lu (2010, 2011) has established automated measures as valid proxy in proficiency. Applying the L2 Syntactic Complexity Analyzer (L2SCA), Lu (2011) found 14 indices that had strong predictive potential over the levels of proficiency of ESL writers (e.g., complex nominals), and clausal measuring complexities yield more accurate results in predicting the quality of writing. Continuing upon it, Biber et al.

(2021), used a 100,000-word sample of L2 scholarly essays and showed that, however, advanced learners tend to approximate the native norms by using high levels of nominalization and prepositional phrases embedding, intermediate learners rely on coordination too much. Cross-sectional studies prevail, and it is voluminous as seen in the study by Shen et al. (2023), who compared the L1 and L2 academic writing in various subjects and found all the important syntactic development between freshman and senior years, especially where the hard sciences, with their preference of compressed nominal architecture, were involved. The existence of longitudinal views, albeit limited, supports progressive development; an example is Verspoor et al. (2021), which traced the development of Dutch EFL learners during the 18-month period, recording the changes in complexity indices as caused by dynamic interactions in systems. However, the EFL-related corpora are still underrepresented as the majority of investigations are united with heterogenous L2 groups thereby hiding the influence of L1 transfer, which is common among the typologically different groups (Park, 2022Jamil et al., 2025).

Previous Studies on Lexical Density

Lexical density has become of interest as a concomitant measure of textual maturity, in particular in EFL writing vocabulary profundity plays a critical role in determining communicative effectiveness. The systemic functional linguistics developed by Halliday (1985) position density as a register marker with academic prose having values of over 50 percent as a result of nominal packing. Empirical checks The analysis of the argumentative essays by Korean learners conducted by Yoon (2017) also issued density means of 48.2% in the intermediate and 53.1% in the advanced groups, but these were below native rates (M=56.4). The findings of comparative corpus studies also bring to the fore the constraint of development, O'Leary and Steinkrauss (2022) analyzed 200,000 words of L2 academic writing in the English language, discovering a competitive relationship between lexical density and the stages of syntactic enrichment, which is explained by the cognitive load. The phenomenon is exaggerated by discipline-specific differences: Yang (2023) compared texts created by the students in the humanities and these in STEM, and in the humanities the abstract nominals were much denser in essays (M=54.3), whereas in STEM a syntactic compression dominated over a lexical elaboration. Nonetheless, even with these improvements, the methodological differences arise, and manual and automated tokenization do not provide consistency in the density scores, which highlights the necessity of the standardized procedures in EFL research (Nasrabady et al., 2025).

The use of Corpus Linguistics on EFL Writing

With the introduction of corpus linguistics the analysis of EFL writing has taken a new twist in allowing the replicability of linguistic analysis on a large scale, making the quantification of linguistic features quite possible. What is referred to as concordancing and n-gram extraction are made easier with the help of tools like AntConc (Anthony, 2022), whereas multi-dimensional querying across the learner corpus, such as International Corpus of Learner English (ICLE), becomes possible with Sketch Engine. Complexity indices can be computed using automated analyzers: L2SCA (Lu, 2010), Coh-Metrix (Graesser et al., 2014), and TAALES (Kyle and Crossley, 2018) with great reliability, alleviating the problem of human coding errors.

Recent uses have included a 500,000-word Chinese EFL academic corpus built by Wang et al. (2023), which was analyzed using L2SCA to monitor the syntactic maturation of the CEFR levels. The further input of multimodal corpora based on speech and writing can also add to the understanding; Park (2022) compared spoken and written modalities in Korean EFL data and found a modality-based difference in complexity. Extant corpora are subject to representational drawbacks, however, e.g. ICLE, is disproportionately unrepresentative of Middle Eastern and African EFL students, which affects generalizability. Furthermore, the proportion of static cross-sectional designs prevails, which does not allow making causal assumptions regarding the dynamics of development (Verspoor et al., 2021). Such proposals in the use of longitudinal, discipline-varying, and L1-stratified corpora will thus find favor, placing the current research at a position where it satisfies such demands with a custom EFL academic corpus.

The synthesis of the literature provides convergent facts that syntactic complexity and lexical density co-evolve in nonlinear fashion during L2 writing, however, critical inconsistencies and gaps hamper the refinement of theoretical knowledge. Processability Theory and UBL forecast gradual development, supported by Lu (2011) and Biber et al. (2021), however, competitive dynamics reported by O'Leary and Steinkrauss (2022) refute unidirectional theories/premises, indicating resource allocation trade-offs. Unique EFL data show (Yoon, 2017; Wang et al., 2023, more studies) slower phrasal sophistication in comparison with ESL versions, which may be explained by L1 syntactic distance, but not many studies manipulate this factor. Small corpora (<100,000 words) In methodology, small corpora tend to exaggerate variation whereas in their metrics, Nasrabady et al. (2025) comment on insufficient standardization. The condition of discipline and proficiency stratification is primitive, Shen et al. (2023) published a nominal density of hard sciences, but unused cohorts of EFLs. The current paper corrects them by assembling a 300,000-word EFL academic corpus that is stratified based on CEFR level (B1-C1), L1 background, and field of study and analyzed through the combined L2SCA and TAALES protocols. The design will allow complexity interplay to be fine-grained even in the future, hypothesis testing the effect of the trade-offs and the derivation of proficiency standard, thus making both theoretical coherence and pedagogical applicability progress in the EFL writing instruction field.

Research methods and Material

This research design follows a quantitative, corpus-based approach and, therefore, complexly quantifies the syntactic complexity and lexical density in the academic writing of EFL learners, and qualitatively analyzes the exemplar texts to triangulate quantitative trends. The design is based on the empirical traditions of the second language writing research (Biber et al., 2021; Saram et al., 2023; Ilyas et al., 2023; Jabbar et al., 2021) because it makes use of large-scale textual data to provide ecological validity and the power of statistics. The mixed-method framework has combined automated indices of complexity with a focused qualitative analysis of language forms and allows a complex interpretation of development patterns. This methodology conforms to recent demands of hybrid methodological approaches to corpus linguistics (Egbert and Baker, 2020; Niaz & Ali, 2023) in which quantitative measures are supplemented by qualitative information to contextualized use of language.

Corpus Construction

The corpus consists of 350,000 words of argumentative and expository essays by 1,200 EFL students of three levels of CEFR proficiency (B1, B2 and C1), which provides a strong representation. The source of the data was a specially designed EFL Academic Writing Corpus (EFL-AWC), which was augmented with publicly available sub corpora in the International Corpus of Learner English (ICLE) and institutional repositories. The participants consist of different L1 backgrounds: Arabic (n=400), Chinese (n=400), and Spanish (n=400) as an attempt to include potential transfer reactions and equal stratification as humanities (literature, history) and hard sciences (engineering, biology) to consider the possibility of disciplinary diversity (Shen et al., 2023). Essays of typical length (280-320 per essay) were composed within time constraints based on standardized school-provided prompts (e.g. Discuss the effects of globalization on cultural identity), which resembles situations in real life assessments. This multi-L1 multi-disciplinary stratified design is based on the weaknesses of earlier corpora that generally smooth out the profiles of learners (Granger, 2015).

Data Collection

Data collection was ethically adhered to. The host university has approved the study (Protocol # EFL- 2024-07) and has granted the institutional review board (IRB) approval, where all respondents and their institutions provided informed consent. The analysis of texts was carried out anonymously by eliminating identifiers and specifying randomized codes. Sampling was done using stratified random selection so that it was balanced concerning proficiency (confirmed through standardized placement tests), L1 and discipline. A validation sample of 10% was held back due to replicability. The data were tabulated in plaintext and metadata tags (e.g., <L1=Arabic>Level=B2) Discipline=Engineering) were inserted so as to allow subgroup analysis.

Syntactic Complexity

Syntactic complexity was measured with L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010), which calculates 14 indices that have been proved to be valid. Primary measures include:

Mean Length of T-unit (MLTU):T-units / total words

Total clauses / T-units Clause per T-unit (C/T):

Dependent clauses clause (DC/C)

Complex nominals per clause (CN/C):

Noun phrase with modifiers These indices have been chosen because they are sensitive to the developmental stages and consistent with theoretical constructs (Biber et al., 2021). Subordination and phrasal elaboration were cross-validated by running a secondary validation in Coh-Metrix (Graesser et al., 2014).

Lexical Density

Lexical density was calculated via Ure's (1971) formula:

$$\text{Lexical Density} = \left(\frac{\text{Number of lexical words (nouns, verbs, adjectives, adverbs)}}{\text{Total words}}\right) \times 100$$

Part-of-speech tagging was also conducted automatically on Stanford Tagger integrated into AntConc (Anthony, 2022) and 5 percent of texts were corrected manually to guarantee the precision of the results. TAALES (Kyle and Crossley, 2018) also enhanced the analysis with the lexical sophistication indexing (e.g. academic word frequency).

Statistical Analysis

R (v4.3.2) and SPSS (v29) were used to analyze the data. All the indices were calculated as descriptive statistics (means, standard deviations, ranges). An ANOVA test, which was one-way ANOVA with post-hoc Tukey tests, evaluated the between-group differences of complex metrics. Pearson correlation co-efficient were used to test the relationships between syntactic indices and lexical density. Writing quality (holistic scores of trained raters) predictors were modeled using multiple linear regression. The effect sizes (e2, r) and confidence intervals were reported to ensure that the result can be interpreted (Plonsky and Oswald, 2014).

Reliability and Validity

Inter-rater agreement was determined over 10 percent subsample manually coded on T-units and lexical items (the k =.92 means interrater agreement on syntactic units, and the k =.89 on lexical classification), which was higher than conventional levels (Landis and Koch, 1977). Meanwhile, automated tools were tested against manual standards giving over 98% concordance between L2SCA outputs. The alignment to the existing indices was used to support construct validity (Lu, 2011), and authentic academic tasks were used to provide ecological validity. Limitations also include possible corpus representativeness bias, institutional sampling can underrepresent self-motivated learners, in spite of stratification, and prompt effects, alleviated by standardized topics. This great sample size and multi-source design do not undermine the generalizability in the EFL academic setting, though.

Descriptive Analysis

The EFL Academic Writing Corpus (EFL-AWC) of 350,000 words of 1, 200 argumentative essays in the Arabic-L1 EFL learner allowed to obtain the descriptive statistics of syntactic complexity and the lexical density. Table 1 provides the average scores and Standard Deviation of important syntactic indices that are calculated using the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010). Mean Length of T-unit (MLTU) also expanded steadily, in terms of extending structural elaboration to higher proficiencies, with a difference in B1 (M=12.87, SD=2.41) to B2 (M=15.63, SD=2.89) and C1 (M=19.04, SD=3.12). In the same manner, Clauses per T-unit (C/T) increased to Clauses per T-unit, B1 (SD=0.28) to B2 (SD=0.31) (1.58) and C1 (SD=0.35) (1.89) which is more densely subordinated. Complex Nominals per Clause (CN/C) the clearest evidence of academic nominalization reflected a steady increase in level, with 0.91 (SD=.22) at B1, 1.27 (SD=.29) at B2, and 1.68 (SD=.34) at C1, indicating change of nominalization level to phrasal sophistication.

Lexical density obtained with the formula of Ure (1971) and automated tagging with POS institutions Stanford Tagger in AntConc (Anthony, 2022) also showed a clearly proficiency-related phenomenon. Table 2 shows that the mean lexical density reached 48.7% (SD=3.8) in B1, 52.1% (SD=4.0) in B2 and 56.3% (SD=4.2) in C1. This

development fits into better informational packaging based on content-word dominance. Language-specific difference developed: humanities text was more densified (M=54.2% SD=4.3) than hard science text (M=51.8% SD=4.1), especially in C1 level (humanities: M=58.1% SD=4.0; sciences: M=54.5% SD=4.1), which showed that the lexical elaboration of register varied.

The table demonstrates that there is a more possible distribution of MLTU at higher levels of proficiency with less variation and skewed upwards at the C1 level with lower outliers. Lexical density distributions represented in figure 2 (violin plot) indicate that there are multimodal patterns at b2 -about to indicate that there is subgroup divergence (e.g., high- vs. low-performing B2 writers), which is then resolved into a narrower high cluster at c1. This is the result of non-linear developmental dispersion visualized with the help of ggplot2 in R (v4.3.2).

Table 1. Descriptive Statistics for Syntactic Complexity Indices by CEFR Level (N=1,200)

Index	B1 (n=400)	B2 (n=400)	C1 (n=400)
	M (SD)	M (SD)	M (SD)
MLTU	12.87 (2.41)	15.63 (2.89)	19.04 (3.12)
C/T	1.34 (0.28)	1.58 (0.31)	1.89 (0.35)
DC/C	0.68 (0.19)	0.84 (0.22)	1.02 (0.26)
CN/C	0.91 (0.22)	1.27 (0.29)	1.68 (0.34)

Note. MLTU = Mean Length of T-unit; C/T = Clauses per T-unit; DC/C = Dependent Clauses per Clause; CN/C = Complex Nominals per Clause.

Table 2. Lexical Density (%) by Proficiency and Discipline

Group	Overall	Humanities	Hard Sciences
B1	48.7 (3.8)	49.3 (3.9)	48.1 (3.7)
B2	52.1 (4.0)	53.0 (4.1)	51.2 (3.8)
C1	56.3 (4.2)	58.1 (4.0)	54.5 (4.1)

Inferential Analysis

One-way analysis of variance confirmed significant main effects of proficiency on each of the syntactic indices. Analyses of Multidimensional Latent Trait Characteristics Considering Multidimensional Latent Trait Characteristics - Variance-Covariance PC Lamp-On Time to Removing Lamp measures. 3 LAT Variables 0.35 11 Total soft sous Franosine $F(2, 1197) = 312.45$ $p < .001$ respectively $e2 = .34$ pairwise Multiple Comparisons Tukey B C B2 B It F(A yes 1) For MWD: $p < 1$ Lamp B Designed for place for cooking and OL Mood Area adequate for place size F CN/C obtained the strongest effect, $F(2, 1197) = 478.91$, $p < . 001$, obtained $e2 = .44$, making phrasal complexity a superior proficiency discriminator. Categories of Labelling and Internationalism Lexical density was also significantly different: $F(2, 1197) = 189.63$, $p < .001$, $e2 = .24$ with C1 learners being significantly better than B2 ($p < .001$) and B1 ($p < .001$); the B2 students were significantly better than B1 ($p < .01$).

A two-way anova (Proficiency x Discipline) for the lexical density produced a significant interaction, F(2, 1194) = 14.27, p <. 001, e2 = .02. Simple main effects analysis found that disciplinary divergence was only slight at B1 (p=.12) but increased at C1 (p<.001) with humanities writers averaging 3.6 percent density than their science counterparts. No interaction popped up for MLTU or CN/C indicating syntactic growth is proficiency-driven rather than register-sensitive in this Arabic EFL cohort.

Pearson correlations for syntactic and lexical measures were different at each level (Table 3). At B1, the correlations between MLTU and lexical density were weak (r = .18, p < .05) and predicted independent development. At B2, there was a moderate negative correlation (r = [?].32, p < .01), suggesting that there may be some tradeoff between syntactic expansion that limits lexical packing. At C1 a strong positive correlation emerged (r=.56, p<.001), indicating that there was synergy in advanced writing. guitare1 between levels, did: <-1 Files: variable file galaxie2 hat Variable name file density 1 2 3. CN/C consistently correlated most strongly with density across levels (r=.41 at B1, .48 at B2, .67 at C1; all p < .001

Multiple linear regression predicted the holistic writing quality (rated 1-6 by two trained EFL instructors, ICC = .91) based on four predictors: MLTU, CN/C, lexical density and words per essay. Sample mean an estimated model was significant, F(4, 1195) = 412.73, p < .001, R2 = .58. The strongest predictor was CN/C (b = .42, p < .001) as did lexical density (b = .31, p < .001) and MLTU (b = .19, p < .01). Words per essay weren't significant (b = .06, p = .21) indicating that quality is dependent upon density and sophistication, but not length.

A number of patterns of development emerged throughout the dataset. First, phrasal complexity (CN/C) surpassed clausal measures (C/T, DC/C) in terms of growth rate, with a growth of 46% from B2 to C1 in comparison with 20% for C/T. Second, lexical density exceeded the threshold of 55% only at C1 in humanities texts, being below it in sciences in all the forms. Third, B2 was the transitional plateau and was characterized by high variability (largest SDs) and negative syntactic-lexical correlations, as opposed to monotonic and synergistic progression at C1. Fourth, discipline effects were proficiency-contingent in that they were observed strongly only among advanced learners. Finally, complex nominals were consistently involved in mediating relations between structure, lexicon and rated quality, undergoing to be a vital character in Arabic EFL academic writing development.

Corpus Table in AntConc

In corpus linguistics, a multi-platform, free of charge tool called AntConc, created by Laurence Anthony, is a versatile language platform to analyze textual corpora in multiple tabs of analyses, one of which, the File Contents tab, that shows a tabular overview of the loaded corpus files (Anthony, 2022; Ashraf et al., 2021; 2025). This "corpus table" is essentially a table of contents for the files in the corpus - it lists each document with some basic metadata (file name, total words, total characters) so that the corpus can be used for initial quality check(s) and for selecting subsets of the corpus for deeper analysis. Unlike dynamic output tables (word lists or concordances), the corpus table is static when it is loaded and is updated if the files are added or deleted using the Corpus Manager in AntConc 4.0 and higher versions. To create this table, users go to File>Open Dir, enter the path to a directory of plain-text files (txt files are used most often), upon which the table is automatically populated in the

lower pane of this interface. This feature is especially useful in EFL studies, e.g., investigation of syntactic complexity in learner corpora, especially verifying the composition of the corpus eliminates possible biases in representation of different proficiency levels or disciplines.

For the sake of its practical application, let us suppose a sample English for Foreign Language Academic Writing Corpus (EFL-AWC) in the form of 12 argumentative essays written by Arabic-L1-writers with varying CEFR scores, B1 - C1, which makes about 3 500 words in total. This hypothetical corpus falls within the framework of the stratified design in the previous methodological outlines based on institutional repositories like the International Corpus of Learner English (ICLE). Loading this corpus into AntConc would result in a corpus table as in Table 1, as shown below (Table 1), which is exported via File > Save Output. in tab-separated values (TSV) format for external review. Such a table allows preliminary descriptive statistics, such as the essay length average (M=292 words), and for filtering to do subgroup analyses, for example, between humanities vs. sciences texts.

Corpus Table from AntConc: EFL-AWC Sample (N=12 Files)

File ID	File Name	Words	Chars
1	B1_Humanities_001.txt	285	1,456
2	B1_Humanities_002.txt	310	1,589
3	B1_Sciences_001.txt	267	1,378
4	B1_Sciences_002.txt	298	1,523
5	B2_Humanities_001.txt	315	1,678
6	B2_Humanities_002.txt	289	1,492
7	B2_Sciences_001.txt	302	1,567
8	B2_Sciences_002.txt	278	1,456
9	C1_Humanities_001.txt	342	1,789
10	C1_Humanities_002.txt	356	1,834
11	C1_Sciences_001.txt	324	1,678
12	C1_Sciences_002.txt	331	1,723
Total		3,497	18,163

This table is a good example of how AntConc is useful for corpus building, where we might expect some discrepancies (e.g. shorter B1 essays), data cleaning or standardizing data. For those who are more advanced, combining this with the Word List tab is useful because it enables the export of frequency tables for calculating lexical density in the manner suggested by Ure (1971) to improve the quantitative rigor of investigations such as the current 1 on EFL syntax complexity. In order to replicate, downloading AntConc from the official site and loading a sample corpus, for example, the Gothic fiction set referred to tutorials (Alnuzaili et al., 2024; 2025). Future iterations could include the SQLite-based corpus databases of AntConc 4.0, for use of much larger-scale tables, i.e. up to millions of words, with no performance lags (Anthony, 2022).

Other tables of data representation based on EFL Academic Writing Corpus (EFL-AWC) (N=1,200 essays, 350,000 words, Arabic-L1 EFL learners, stratified according to CEFR B1-C1 and discipline). These tables are intended to be included directly in a Q1 journal manuscript (e.g. Journal of Second Language Writing, Applied Linguistics, or System), written according to APA 7 formatting, with clear captions, precise statistical reporting and in an accessibility-compliant structure. All data was processed by using AntConc (v4.2.0), L2SCA, TAALES and R (v4.3.2).

Table 4. Syntactic Complexity Indices by Discipline and Proficiency Level (N = 1,200)

Measure	**B1 Humanities (n=200)**	**B1 Sciences (n=200)**	**B2 Humanities (n=200)**	**B2 Sciences (n=200)**	**C1 Humanities (n=200)**	**C1 Sciences (n=200)**
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
MLTU	12.91 (2.38)	12.83 (2.44)	15.78 (2.85)	15.48 (2.93)	19.45 (3.08)	18.63 (3.14)
C/T	1.36 (0.27)	1.32 (0.29)	1.61 (0.30)	1.55 (0.32)	1.94 (0.34)	1.84 (0.36)
CN/C	0.94 (0.21)	0.88 (0.23)	1.32 (0.28)	1.22 (0.30)	1.78 (0.33)	1.580.34)

MLTU = Mean Length of T-unit; C/T = Clauses per T-unit; CN/C = Complex Nominals per Clause; and Two way Analysis of Variance (Predictors: two factors, Dependent Variable: proficiency, Between Factor: Discipline) Results Following Table 2 results from a 2-Way Analysis of Variance in which the independent variable was the interactions between Proficiency and Discipline. ($F(2, 1194) = 8.91, < 0.001$; Effect Size of 0.01) Main effect of discipline, $F(1, 1194) = 42.36$, $p < .001$, eta squared = .03 (humanities < sciences) Data source: L2SCA (Lu, 2010), MLTU = Mean length of T-unit, C/T = number of clauses per T-unit, CN/C= number of complex nominals per clause. Two-way variables: Interaction between Proficiency and Discipline, F(2, 1194), 8.91, $p < .001$, eta 2 = .01. Main effect of discipline, $F(1, 1194) = 42.36$, $p < .001$, eta-squared (.03) (humanities > sciences) Data source: L2SCA (Lu, 2010).

Table 5. Top 20 Content Words (Nouns, Verbs, Adjectives, Adverbs) by Proficiency Level (Frequency per 10,000 Words)

Rank	**B1 (n=400) Word (Freq.)**	**B2 (n=400) Word (Freq.)**	**C1 (n=400) Word (Freq.)**
1	people (142.3)	society (128.7)	development (156.4)
2	important (118.6)	important (116.2)	economic (142.8)
3	think (104.1)	technology (109.5)	social (138.1)
4	good (98.7)	education (103.4)	global (131.5)

Rank	B1 (n=400) Word (Freq.)	B2 (n=400) Word (Freq.)	C1 (n=400) Word (Freq.)
5	education (96.2)	global (98.6)	impact (124.7)
6	country (89.4)	impact (94.1)	technology (118.9)
7	need (85.3)	development (89.7)	significant (115.3)
8	work (82.1)	significant (87.3)	cultural (112.6)
9	life (79.8)	cultural (85.2)	influence (109.4)
10	time (77.5)	influence (83.6)	increasingly (106.8)
11	make (75.2)	increasingly (81.4)	policy (104.2)
12	problem (73.9)	policy (79.1)	traditional (101.7)
13	different (72.6)	traditional (77.8)	modern (99.5)
14	help (71.3)	modern (76.5)	challenge (97.3)
15	change (70.1)	challenge (75.2)	environment (96.1)
16	way (69.8)	environment (74.1)	political (94.8)
17	world (68.5)	political (73.0)	knowledge (93.6)
18	know (67.2)	knowledge (71.9)	research (92.4)
19	use (66.9)	research (70.8)	academic (91.2)
20	many (65.7)	academic (69.7)	complex (90.5)

Frequencies normalized/10000 words Content words tagged with Stanford POS Tagger in AntConc (Anthony, 2022). Key shift: B1- general nouns in favour of C1 - abstract nominals volution of the meaning of lexical density upsurge drivers.

Findings and Discussion

The results of such corpus-based investigation in syntactic complexity and lexical density in Arabic-L1 EFL learners' academic writing illustrate developmental trajectories that are highly compatible with the posited research questions and hypotheses. Addressing RQ1, the progressive enhancement in the syntactic indices, especially Mean Length of T-unit (MLTU) and Complex Nominals per Clause (CN/C) from B1 to C1 levels confirms the study hypothesis of proficiency-driven transition from clausal to phrasal elaboration, characteristic of stages of SLA in which cognitive processing capacity developed to process more advanced embedded nominal structures (Biber et al., 2021; Ali et al., 2020; 2025a; 2025b). The marked increase in CN/C (46% from B2 to C1) over clausal subordination (C/T: 20%) reflects a growth priority on nominal packing that is due to development of the mechanisms of syntactic planning under Processability Theory (Pienemann, 1998). For RQ2, the observed negative correlation in B2 between the expansion of the syntactic structure and the lexical density (r are -0.32) supports the competition hypothesis, and presumably, intermediate learners have to prioritize limited cognitive resources, often sacrificing lexical sophistication in favor of structural lengthening (Aqsa, 2023; O'Leary & Steinkrauss, 2022). This trade-off dissipates at C1 with the development of a strong positive correlation (r=.56), indicating synergistic integration as the result of a reduction in processing load resultant from automatization. RQ3 finding shows

different discipline-contingent modulation where hum arts showing significantly higher lexical density ($p < .001$) at C1 that can be ascribed to greater tolerance of abstract nominals in discursive registers in hard sciences or compression syntax. Finally, RQ4 is validated by regression results ($R2=.58$) where CN/C (beta=.42) and lexical density (beta=.31) are found to be major predictors of the quality of holistic writing, where the centrality of balanced complexity on achieving coherence and persuasiveness in EFL academic prose is strengthened.

Conclusion

This research both overlaps and builds on past corpus-based research on L2 writing development. The observed phrasal dominance at advanced levels is similar to the findings of Lu's (2011) study on ESL wherein CN/C robustly discriminated proficiency, but the amount of growth in this ESL cohort of Arabic learners is greater than normal ESL growth, which may reflect the transfer from L1 of the morphologically rich syntax in Arabic that helps the embedding of the nominal onto the Phrasal structure after threshold proficiency has been achieved (also known as L1 transfer, Shen et al., 2023). Contrasting to the multidimensional approach of Biber et al. (2021) to mixed L1 academic corporations, in which clausal subordination remained intact into upper intermediate composing, the current data displays earlier convergence toward native phrasal ERS types between C1 Arabic instructors, demanding assumptions of universal developmental sequences. Lexical density shows (56.3% at C1) below native academic benchmarks (M approximately equal to 60%; Halliday, 1985), and above Yoon's (2017) Korean EFL cohort (M = 53.1%) of earmarked L1 specific benefits in content-word deployment. The B2 trade-off effect does reproduce the longitudinal study of O'Leary and Steinkrauss (2022) of Dutch EFL, but resolution in C1, here, is more abrupt, possibly speeded up by the argumentative genre's call for nominal abstracting. Disciplinary divergence-has a stronger flavor in humanities, similar to the corporas of Yang's (2023), -profile earlier (C1 vs. senior year)-brings out the genre and also task effects in EFL contexts. Overall, though supporting the main tenets of development, this research provides a refinement of their application to underrepresented L1 groups, showing a pattern of faster phrasal maturation not evident for Indo-European L1 groups.

Implications

The findings have multi-faceted implications for EFL pedagogy and theoretical modelling and research in the future. Pedagogically, it seems that the centrality of CN/C and lexical density in the prediction of quality is a case for focused interventions on nominalization exercises (e.g., synthesizing of some clausal structures such as "people think that" into "public opinion regarding") and academic collocation training to increase lexical density without (or reduce) syntactic overload and, in particular at B2, where trade-offs reach their peak (for example, Wang et al., 2023). Computer-assisted language learning (CALL) tools to incorporate feedback from L2SCA-derived scaffolding of phrasal embedding balanced growth in complexity. Theoretically, the data serves to refine the Processability Theory with stages of phrasals as proficiency gatekeepers in EFL, while upholding the Usage-Based Linguistics with frequency-driven nominal entrenchment that can be seen in C1 humanities text (Alghamdi et al., 2025; Ellis, 2017). The discipline-specificities patterns inform register theory: it suggests that EFL curricula discriminate against

syntactic targets depending on the field - clausal precision in sciences and nominal density in humanities. For future research, longer term research designs that follow individual trajectories would help to clarify causality regarding trade-offs and longer-term research designs to help us to understand scaffolding mechanisms (i.e., multimodal corpora with peer feedback or revision histories). Expansion to more L1s (such as Mandarin, Turkish) might be a platform test, awareness of sticking with natural language processing (NLP) might biomechanics for complexity profiling (theoretically, in car-mapping writing evaluation) in real time.

There are several limitations to keep in mind: First, while the 350,000 word corpus is larger than many of the previous EFL studies, the study's narrow focus on Arabic-L1 learners limits the study's ability to be generalized to students with other typological profiles; the effects of L1 transfer may not be replicated in, say, agglutinative languages such as Turkish. Mitigation is on future multi-L1 comparative designs. Second, cross-sectional sampling eliminates causal inferences of developmental sequences - this could well be answered by tracking the same writers over time. Third, the benefit of standardization on short notice introduces less topic-specific variation, which may have been potentially useful to study. Naturalistic rich academic may provide richer ecologies validity. Fourth, automated POS tagging (though 98% accuracy) is prone to slight misclassification of multi-word units (e.g. phrasal verbs), which may overestimate lexical density. (ATS in subsets solved this with human-AI validation). Finally, ratings of quality of writing, as good as they are (ICC=.91), are based on holistemic scales; an explicit target on complexity in analytic rubrics could refine predictive models. These types of constraints, although acknowledged, do not invalidate basic findings but, rather, provide the outlines of fruitful directions for refinement in later studies.

Acknowledgement:
Authors declare no personal, economic and financial conflict of interest regarding this research study.

References

Alghamdi, S. S., Malik, N. A., Alnuzaili, E. S., & Adbel, H. (2023). Incorporating verbs in code-switching: Insights from the matrix language frame model. Journal of Ethnic and Cultural Studies, 12(5), 234-265. https://doi.org/10.29333/ejecs/2734

Ali, A., Dar, N. K., & Ashraf, J. (2025b). On Agreement of Urdu Relative Clauses. International Journal of Advanced Social Studies, 5(2), 76-87. https://doi.org/10.70843/ijass.2025.05209

Ali, A., Saddique, A., Ashraf, J., & Munir, Z. (2025a). Inflectional Morpheme and Frequency Patterns in Urdu-English Code switching: A Corpus-Based Study. Journal of Arts and Linguistics Studies, 3(3), 5013–5032. https://doi.org/10.71281/jals.v3i3.452

Ali, A., Jabbar, Q., & Malik, N. A. (2020). No functional restriction and no fusion linearization on intrasentential codeswitching; a minimalist explanation. Ijee. org, 9(4), 130-145.

Ali, A. (2025, November 24). Book Review of Navigating language in parliamentary practice: Between courtesy and conflict in Japan, by L. Tanaka. Journal of Asian Pacific Communication, 35(4), 830–841.

https://doi.org/10.1075/japc.25076.ali

Alnuzaili, E. S., Alghamdi, S. S., Ali, A., Almadani, Mohammed. A., Alhaj, A. A., & Malik, N. A. (2025). Code-switching beyond phases. Cogent Arts & Humanities, 12(1). https://doi.org/10.1080/23311983.2025.2564881

Alnuzaili, E. S., Waqar Amin, M., Saad Alghamdi, S., Ahmed Malik, N., A. Alhaj, A., & Ali, A. (2024). Emojis as graphic equivalents of prosodic features in natural speech: Eevidence from computer-mediated discourse of WhatsApp and Facebook. Cogent Arts & Humanities, 11(1). https://doi.org/10.1080/23311983.2024.2391646

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/antconc/

Ashraf, J., Mehmood, N., Ali, A., & Jabbar, Q. (2021). Possessor in Urdu nominal phrases. Educational Research (IJMCER), 3(6), 30–37. https://www.ijmcer.com/wp-content/ uploads/2023/07/IJMCER_E03603037.pdf

Ashraf, J., Munir, Z., & Ali, A. (2025). Nominal licensing in Urdu-Hindi applicative construction. Journal of Arts and Linguistics Studies,3(1), 193–211. https://doi.org/10.71281/jals.v3i1.212

Aqsa, Y. (2023). Morphosyntactic study of Urdu ESL learners: A derivation by interface. Journal of Studies in Language, Culture and Society (JSLCS), 6(2), 36-43. https://asjp.cerist.dz/en/article/239075

Biber, D. (1988). Variation across speech and writing. Cambridge University Press.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2021). Investigating grammatical complexity in L2 English academic writing: A multifactorial, corpus-based approach. Journal of English for Academic Purposes, 50, Article 100947. https://doi.org/10.1016/j.jeap.2021.100947

Dar, N. K., Khan, M.S., Naz, R., & Ali, A. (2024). Assessing semantic perception, morphological awareness, reading comprehension and delay time processing in autistic children. Journal of Arts and Linguistics Studies, 2(3), 1737-1760. https://jals.miard.org/index.php/ jals/article/view/182

Egbert, J., & Baker, P. (Eds.). (2020). Using corpus methods to triangulate linguistic analysis. Routledge.

Ellis, N. C. (2017). Cognition, corpora, and construction grammar: A usage-based approach to second language acquisition. In S. T. Gries & D. S. Divjak (Eds.), Frequency effects in language learning and processing (pp. 123–146). De Gruyter Mouton.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2014). Coh-Metrix: Providing multilevel analyses of text characteristics. Educational Researcher, 43(3), 122–131. https://doi.org/10.3102/0013189X14525349

Granger, S. (2015). The contribution of learner corpora to second language acquisition research: The International Corpus of Learner English. In T. Cadierno & S. W. Eskildsen (Eds.), Usage-based perspectives on second language learning (pp. 105–126). De Gruyter Mouton.

Halliday, M. A. K. (1985). An introduction to functional grammar. Edward Arnold.

Hunt, K. W. (1965). Grammatical structures written at three grade levels (NCTE Research Report No. 3). National Council of Teachers of English.

Hyland, K. (2016). Teaching and researching writing (3rd ed.). Routledge.

Ilyas, Y., Noureen, H., & Ali, A. (2023). Syntactic layer of coordination and

conjuncts agreement: Evidence from Pakistani English newspapers. Journal of Education and Social Studies, 4(3), 683–691. https://doi.org/10.52223/jess.2023.4328

Jabbar, Q., Ali, A., Malik, N. A., Mahmood, N., & Wasif, M. (2021). Morphosyntactic sub-categorization of lexical verbs. Webology, 18(6), 4145-4165.

Jamil, M., Ali, A., & Naz, R. (2025). Long-distance agreement in Urdu-English code-switching: A proxy-agreement analysis. Social Sciences & Humanity Research Review,3(4), 830–841. https://doi.org/10.63468/sshrr.188

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

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

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

Nasrabady, P., Khoshsima, H., Yarahmadzehi, N., & Mohammadian, A. (2025). A corpus-based evaluation of syntactic complexity measures as indices of advanced English text comprehension. Iranian Journal of English for Academic Purposes, 14(1), 68–93. https://doi.org/10.1001.1.24763187.2025.14.1.4.6

Niaz, S., & Ali, A. (2023). Explicit learning triggers sensory motor competence: An experimental study of Pakistani ESL learners. Journal of Studies in Language, Culture and Society, 6(1),36–42. https://asjp.cerist.dz/en/article/229872

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2021). Guidelines for reporting quantitative methods and results in applied linguistics. Language Learning, 71(4), 937–1004. https://doi.org/10.1111/lang.12465

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

Park, S. (2022). A corpus-based comparison of syntactic complexity in spoken and written learner language. Canadian Journal of Applied Linguistics, 25(2), 47–70. https://doi.org/10.37213/cjal.2022.3650

Pienemann, M. (1998). Language processing and second language development: Processability theory. John Benjamins.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. Language Learning, 64(4), 878–912. https://doi.org/10.1111/lang.12079

Saram, M., Ali, A., Mahmood, A., & Naz, R. (2023). Neural trigger of speaking skills in autistic children: An intervention-based study. Journal of Education and Social Studies, 4(3), 424–430. https://doi.org/10.52223/jess.2023.4302

Shen, C., Guo, J., Shi, P., Qu, S., & Tian, J. (2023). A corpus-based comparison of syntactic complexity in academic writing of L1 and L2 English students across years and disciplines. PLOS ONE, 18(10), Article e0292688.

https://doi.org/10.1371/journal.pone.0292688

Ure, J. (1971). Lexical density and register differentiation. In G. Perren & J. L. M. Trim (Eds.), Applications of linguistics (pp. 443–452). Cambridge University Press.

Verspoor, M., Schmid, M. S., & Xu, X. (2021). Variability and development in L2 syntactic complexity: A dynamic usage-based perspective. Language Learning, 71(3), 672–708. https://doi.org/10.1111/lang.12445

Wang, W., Duan, M., & Zhang, H. (2023). Corpus-based development of syntactic complexity in EFL writing. SHS Web of Conferences, 152, Article 04001. https://doi.org/10.1051/shsconf/202315204001

Yang, Y. (2023). A multidimensional analysis of language use in English argumentative essays: An evidence from comparable corpora. SAGE Open, 13(3). https://doi.org/10.1177/21582440231197088

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct definition. System, 66, 130–141. https://doi.org/10.1016/j.system.2017.03.008