ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

ANALYZING LEXICAL COMPLEXITY IN LEARNER CORPORA:A CORPUS-DRIVEN APPROACH USING PART-OF-SPEECH TAGGING AND DEPENDENCY PARSING

Menahil (Corresponding Author)
MS Scholar, Applied Linguistics, National University of Computer and Emerging Sciences, Lahore, Pakistan.
Email address: menahilramzan2001@gmail.com
Saadia Khan
PhD Scholar, English (Linguistics), University of Education, Lahore, Pakistan.
Email address: saadiakhanabdali@gmail.com

Abstract
Lexical complexity is a crucial component of second language (L2) proficiency, encompassing the range, sophistication, and density of vocabulary used by learners. This study examines lexical complexity in learner corpora through a corpus-driven methodology leveraging part-of-speech (POS) tagging and dependency parsing. The objectives are to (a) operationalize lexical complexity into measurable indices (lexical density, diversity, and sophistication), (b) utilize POS tagging to automatically identify and categorize lexical items, and (c) apply dependency parsing to incorporate syntactic context into lexical complexity analysis. The study analyzes a corpus of 300 L2 English essays from intermediate and advanced learners. Results show that advanced learners use a higher proportion of low-frequency "sophisticated" words and exhibit greater lexical diversity than intermediate learners, although lexical density (content word ratio) remains comparable across proficiency levels. Dependency-based metrics (e.g., average dependency length, noun modifier counts) provided additional insights into how learners deploy complex lexico-syntactic structures. The findings highlight a positive correlation between lexical complexity indices and writing quality scores, with lexical sophistication and diversity emerging as significant predictors of human-rated proficiency. The study fills a methodological gap by integrating NLP tools in learner corpus research to yield a multi-dimensional profile of lexical complexity. Implications are discussed for L2 writing pedagogy, automated writing evaluation, and the development of hybrid computational models for L2 complexity. This research underscores the value of combining POS tagging and dependency parsing in corpus analyses to obtain granular, robust measures of lexical complexity in learner language.

Keywords: Lexical complexity; learner corpora; POS tagging; dependency parsing

Introduction

An essential feature of writing proficiency in a second language is accurate grammar in addition to the need for a deep and contextually appropriate vocabulary. Within applied linguistics, the lexicogrammar complexity of learner output has been well recognized as one of the important measures of a language learner's proficiency. In a general sense lexical complexity is the expression of the sophistication, diversity and density of vocabulary used by the learners in the written and oral discourse. However, lexical complexity is difficult to measure and analyses as it is a multi-dimensional phenomenon reflecting a variety of components in lexical knowledge and application. The present study adopts a corpus driven approach to investigate lexical complexity in learner corpora, making use of the latest natural language processing methods (i.e., part-of-speech tagging and dependency parsing) to attain a fine-grained analytical scheme. The introduction provides a background to lexical complexity and its importance, it states the objectives of the study and the research questions which will be answered, and finally it states the problem that the study is addressing.

In second language acquisition (SLA) literature, lexical complexity (also called lexical richness) subsumes a number of sub-constructs, the first of which is lexical diversity (the amount of vocabulary used), the second of which is lexical sophistication (the use of low-frequency or advanced words), and the third of which is lexical density (the proportion of

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

content words used). Taken together, these dimensions represent the scope and spread of a learner's lexical inventory in genuine language use. Previous studies have proven that lexical richness is very closely linked to L2 achievement as well as quality of writing. For instance, high-level L2 learners have a larger number of lexical items (high diversity as well as the high frequency of very low-frequency, academic lexical items high sophistication) as compared to the lower-level learners who have a simpler set of common words. Correspondingly, lexical density (CW/TW) does not necessarily increase with proficiency; it sometimes depends on task type or genre; lower-level learners are sometimes able to produce texts containing a lot of content in certain genres. Overall, lexical complexity is preferred as one of the most important aspects of communicative competency, as a strong vocabulary allows for more accurate and complex communication.

Learner corpus - the body of texts created by L2 learners in electronic format - has become an indispensable resource for exploring the lexical complexity. They give researchers the opportunity to carry out systematic investigations of authentic learner language on a large-scale basis. By analyzing such corpora, patterns can be discovered, for example, of the over or underuse of certain word classes, of the changes in the development of vocabulary over time or of the differences between corpora composed of learners of different proficiency levels or of different L1 backgrounds. The corpus-based studies conducted so far have yielded a lot of information. For example, a recent large-scale study of Chinese EFL university students' writing showed that though their lexical density was similar to that of native English students, their lexical sophistication and diversity were much lower. This means that L2 writers are able to communicate as much information (content words) as native speakers, but they are using a more limited and less sophisticated lexicon. Another corpus study of high-stakes English examinations in Sweden showed that in terms of lexical complexity, they influenced the grading of the essay in the way that instructors and expert raters did not always agree due to differences in their sensitivity to use of advanced vocabulary by students. These findings highlight the pedagogical impact of lexical complexity and the need for accurate analytic approaches to quantify it.

Nonetheless these developments, problems remain related to the best operationalization and analytical strategies for lexical complexity. Conventional metrics, such as the type conservative ratio (TTR) have been shown to suffer from a high degree of sensitivity to the length of the text, which has led to the development of more recent indices (e.g., the measure of text diversity (MTLD), the type conservative ratio under expected frequency distribution (VOCD)), which are advanced to be robust indicators of lexical diversity. Lexical sophistication indices usually use high-class frequency corpora or lexicons to identify the terms that might be placed in the category of "advanced". Moreover, empirical corpus research has shown lexical complexity to be modulated by variables of context such as task type and subject matter. For example, Yoon and Polio (2017) noted that the topic of the essay may influence both lexical and syntactic complexity measures in such a way that the comparability of measures in response to disparate writing prompts may be suspect. These findings highlight the need for multidimensional, context-aware methodological approaches to the study of lexical complexity.

Objectives

1. To develop a hybrid computational model that measures lexical complexity in L2 learner writing using POS tagging and dependency parsing.
2. To analyze how lexical complexity indices relate to human-rated writing quality across learner proficiency levels.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Research Question

1. How can part-of-speech tagging and dependency parsing be applied to analyze lexical complexity in L2 learner corpora, and what differences emerge across proficiency levels?
2. How do automated lexical complexity measures—such as diversity, sophistication, and density—relate to human-rated writing quality when analyzed through a hybrid POS and dependency-based approach?

Problem Statement

Conventional approaches to textual complexity forced to lexical complexity often stick to crude frequency counting and ratio calculation that ignores contextual and syntactic views of vocabulary use. Manual analysis creates constraints in terms of scalability and objectivity and hinders the proper uncovering of nuanced patterns in learner writing. In addition, current measures do not take into account the syntactic interactions that express complexity. This work removes these shortcomings by combining part-of-speech tagging and dependency parsing, to give a more accurate and context sensitive measure of lexical complexity. The purpose is to address the methodological and empirical gaps in assessing how lexical sophistication is an index of learner proficiency.

Literature Review

Research on lexical complexity spans multiple subfields of applied linguistics, including second language writing, vocabulary acquisition, and corpus linguistics. This literature review synthesizes recent findings (2017–2025) in five key areas relevant to our study: (1) definitions and dimensions of lexical complexity in L2 writing, (2) corpus-driven studies of learner lexical complexity, (3) the role of part-of-speech tagging in lexical analysis, (4) the use of dependency parsing in complexity research, and (5) integrated (hybrid) approaches and the relationship of lexical complexity to proficiency and writing quality. Each subsection draws on contemporary research to contextualize our approach and highlight the gap our study aims to fill.

Lexical Complexity in Second Language Writing

Lexical complexity forms a multifaceted construct in the area of second language acquisition (SLA), and comprises a number of vocabulary-related attributes. Recent scholarship emphasizes the fact that no single metric can adequately capture the range of lexical complexity. Lexical diversity describes the range of different lexical items (types) that are used by the learners. While the type/token ratio is often a common parameter to assess this indicator, other sophisticated versions of this statistic have been proposed, such as the Measure of Textual Lexical Diversity (MTLD), to control for the effects of the length of the text. Lexical sophistication, which is also referred to as lexical rarity and/or advanced usage, refers to the percentage of less frequent or academic terms in a discourse. Operationally, this often translates to frequency bands, as the percentage of words that are rare in some reference corpus or that fall into higher frequency bands of a standardized frequency list (e.g., words beyond the 5,000 most prevalent words). The relationship between lexical density and content is that lexical density is the number of content words, which include nouns, lexical verbs, adjectives, and adverbs, over the number of words in the total text which can be seen as the informational density of the text. The triad of diversity, sophistication and density is widely recognized as the basic building blocks of lexical complexity.

Empirical studies in the past decade emphasized how important the lexical complexity is to the quality and growth of L2 writing. For example, Kim, Crossley and Kyle (2018) modelled lexical sophistication as a multidimensional phenomenon and showed that a number of statistically distinct components, relating to frequency, psycholinguistic properties of words

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

(concreteness, imageability), n-gram frequency and word acquisition sequence, composite lexical sophistication. Their analysis of EFL students' project proposals showed that differential aspects of lexical sophistication (for example, usage of the academic versus colloquial vocabulary) were differentially related to instructor assessments of writing quality. This finding highlights the fact that lexical complexity is not a monolithic concept, rather there are a set of different lexical features that contribute to writing proficiency, and the relative importance of these features may vary from context to context or genre to genre. Likewise, Lu and Hu (2022) argued to see lexical sophistication as multifaceted. They mentioned that the former indices hardly ever took into account polysemy (multiple semantic senses of a word) and thus created sense-aware lexical indices to overcome this insufficiency. Their results confirmed that considering the senses of words increased the predictiveness of L2 writing quality, and with that, provides an advanced view of what constitutes "sophisticated" lexical usage. In other words, sophistication is achieved not by just using a difficult word, but by making such a usage in a sophisticated way.

Lexical complexity is also intertwined with syntactic complexity, at times with a compensatory relationship. The Complexity Trade-off Hypothesis is, that languages (and by extension, language performances) may balance their complexity across domains, that is when one aspect of a language, such as syntax, is too complex, another aspect, such as the lexicon, may show a certain relative simplicity, and vice versa. In learner language, this has taken the form of a possible competition between the growth of syntactic structures and lexical growth. O'Leary and Steinkrauss (2022), for example, explored language acquisition pennyweight (L2) academic writing and found evidence of a trade-off, where the more learners focused on the incorporation of advanced vocabulary (lexical bloom), the more (syntactic retreat) complexity of their sentence constructions would be marginally reduced at specific developmental stages. This is not to say that both domains cannot be improved simultaneously, rather it refers to both cognitive constraints and pedagogical practices (e.g., an instructional emphasis on vocabulary may temporarily decrease attention to complex sentence construction). As mentioned, acknowledging this interplay is key, since it suggests that lexical complexity should be explored in connection to syntactic patterns in order to comprehensively evaluate learner performance - one reason for the utilization of dependency parsing in the present study.

Finally, the topic and genre influence should be considered. Yoon and Polio (2017) re-examined the validity of the metrics of lexical complexity by analyzing essays of diverse topics and genres (narrative versus argumentative). They found that specific topics naturally induced a wider range of vocabulary and therefore inflated diversity scores, while technical or specific topics caused the repetitive use of certain terms and thus deflated diversity scores. Genre also played a decisive role: argumentative prose tended to make more use of abstract nouns and academic terminology (i.e., higher sophistication) than personal narratives. These results suggest that researchers need to interpret lexical complexity measures against the context of relevant situational factors because raw numbers can be deceiving - for example, a low diversity score could be due to a narrowly constrained prompt and not a valid reflection of a learner's lexical competence.

Corpus-Driven Approaches to Learner Lexical Complexity

Corpus studies of L2 learners have made an immense contribution to our knowledge of lexical complexity by providing authentic data for quantitative analyses on which close examinations are based. In recent years, there has been a noticeable growth in corpus-based research of the formation of L2 lexical complexity and use of this complexity as a discriminating feature between populations of learners. Most of these studies involve the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

collection of student writing corpora (sometimes longitudinal or with a variety of proficiency levels), and then the computation of a range of complexity measures.

One of the major strengths of corpus-driven methods is their ability to reveal patterns which are invisible under small samples or intuitive analyses. For example, Holmberg Sjöling (2025) worked with a self-constructed corpus of exam essays in order to investigate how lexical complexity affects teacher assessments of the quality of texts written by Swedish EFL learners. Using TAALED and TAALES, which can assess lexical diversity and sophistication of essays in the high stakes' exam, Holmberg Sjolding found that expert raters (the teachers who are trained by the national agency) rewarded lexical complexity more uniformly than classroom teachers did. Teachers tended to be more forgiving or be more tolerant of lexical sophistication absence or were at times blind to its presence, whereas expert raters expected it at a higher level of proficiency. It is clear from this corpus-based result that lexical complexity can influence assessment results and that learners may need assistance not only in expanding their lexical knowledge base, but also in using that lexical knowledge in communication situations to meet the expectations of the raters.

A corpus-based study was also conducted by Yang and He (2025) in which they compared a corpus of Chinese university students' EFL essay with the reference corpus of Native English student essay. The authors stated that Chinese EFL learners had similar lexical density as native Chinese EFL learners, which means they had a similar proportion of content words in their writing. However, on two other dimensions - lexical sophistication (measured by percentage of low-frequency and academic words) and lexical variation (diversity) - learners were significantly ahead of their native writer counterparts. With the data from the corpus, it was revealed that even the advanced EFL learners tend to utilize a basic and common vocabulary and do not have an opportunity to express a richer and advanced linguistic item. Yang and He explained that the similar density might be due to first language effects, with Chinese writing being value-driven with respect to conveying information, and therefore transferring that habit into English writing lessons, feeding as much information into the writing as possible. This paper illustrates the powerful ways that corpus studies can reveal cross-linguistic effects on lexical style.

Corpus research has also shed light on developments over time. Keeping an eye on the same learners over time to use longitudinal corpora, although they are rarer due to the demands, are extremely valuable. For instance, Xu and Casal (2023) used a longitudinal study of L2 legal writing in an English-for-Specific-Purpose context across an academic year. Using dependency-parsed corpora and lexically frequency analysis, they found a distinct trend: learners gradually added more complex legal terminology (lexical sophistication) and over time started to use it in complex collocations, as part of longer noun-skips, and in other syntactic patterns. These results integrate the lexical and the syntactic data and are consistent with insights from Complex Dynamic Systems Theory according to which the subsystems of language grow in harmony with occasional bursts and plateaus. Xu and Casal's study illustrate the usefulness of the so-called parsed corpora in which each word is marked with grammatical relations - an approach which is used in this study.

Apart from learner production corpora, reference corpora and wordlists have been used to measure the lexicogrammar of learner texts. One change that has become prominent since 2017 has been the use of large-scale frequency lists such as the CEFR lexical lists and the New Academic Word List in order to analyze words in learner corpora according to degrees of difficulty. Zhang and Lu (2025) used 24 linguistic features, most of which are lexical by nature (such as the distribution of words in a CEFR book) to train a text difficultreading model for L2 readers. Their corpus-based model used to assign bands to the texts was accurate with

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

approximately 82.6% success at broad level distinctions, confirming the strength of systematically extracted lexical features from corpora as predictors of text complexity. Such studies demonstrate the benefits of corpus data and computational models for informing the selection of material to learners so that reading texts are lexically matched to learners' ability levels.

Part-of-Speech Tagging in Lexical Complexity Analysis

Part-of-speech (POS) tagging is the process of automatically assigning grammatical categories (e.g., noun, verb, adjective, etc.) to each lexical item in a text. In corpus linguistics, POS tagging is a basic pre-analytic stage that enables fine-grained investigations of lexical structure. For studies of lexical complexity, POS tagging affords at least two important benefits: (1) it makes it possible to distinguish content words from function words, and (2) it opens up the possibility of calculating POS-specific lexical indices, such as measures of noun diversity or adjective frequency. Both of these functions are important as they make evaluations of lexical richness more accurate and informative.

A main use of POS tagging is the lexical density estimation. Lexical density, calculated as the share of content words in a text, can only be calculated accurately if all the tokens are always classified. While manual tagging of content words is possible for small samples, large corpora cannot be tagged by a human annotator, because classification will not be applied consistently to all words. Empirical studies using POS tagging yield lexical density statistics that can be compared across learners. For instance, Yang and He [2025] used POS-tagged corpora to show that lexical density does not vary significantly between Chinese EFL learners and native English speakers, and their tagger found content words in each essay to make such comparison possible at scale. Moreover, tagging showed that the Chinese learners used similar proportions of nouns, but less of the unique adjectives and adverbs, which impacted on the comparative qualitative nature of the text. Some of these findings, such as the fact that learners' tendency to use main verbs at the expense of light usage of modal and auxiliary verbs is apparent in their tendency to use nouns and few modifiers are directly attributable to POS-marking.

Besides density, POS tagging can be used to calculate lexical diversity in discrete grammatical categories. A learner can show a wide variety of nouns and a limited use of verbs and vice versa. In addition, relative lexical composition may mask overall diversity; diversity tags can be applied to each POS and diversity per POS can be examined to build a detailed lexical profile. One study found that advanced learners showed a rich range of verbs whilst using a rather generic choice of nouns (e.g., 'thing', 'person', 'problem'). This kind of information is used in instructional design to show where more attention needs to be made in the classroom to accurate nouns or descriptive adjectives. In TAALES 2.0, Kyle et al (2018) added POS-specific frequency indices (e.g., average frequency of noun vs. verbs) to the 316 analytical variables. These indices allow researchers to know, for example, whether a learner's usage of verbs is biased towards lower-frequency lexis (which is an indicator of more advanced lexical maturity), while his or her choice of nouns is still characterized by high-frequency forms. The ability to break down sophistication by POS was a major methodological breakthrough for such tools.

Another area in which POS tagging has a lot of applications has to do with finding academic vocabulary in context. Most academic words can be a nominalization or adjective in form (e.g., implementation, significant). By annotating a corpus, the study can precisely determine the percentage of academic nouns used. North and Zampieri (2023) undergoing a study on perceptions of lexical complexity highlighted the importance of POS by analyzing word length, frequency, and syllable count. Although word level features (word length, the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

number of syllables, etc.) were used, their study relied on accurate POS tagging in order to interpret these features properly (e.g., the long token is not a hard proper noun, but a hard phrasal verb). Their results further showed that word frequency and word length had a stronger impact on rated lexical complexity for L2 learners than for L1 learners. A full-fledged POS tagger would make it possible, then, to investigate in detail whether indeed such perceptions differ systematically between word types, for example whether long verbs are going to be particularly challenging for L2 speakers.

In our particular context, the use of POS tagging of the learner corpus allows us to introduce one important control: the study consider the measures of lexical diversity and sophistication based on content words only. Determiners, prepositions, pronouns and simple conjunctions are part of the closed class of function words, and are actually very commonly used, even by elementary students (e.g., and, the, to, of). Counting as diversity can give falsely high figures for apparent diversity; a text rich in function words can superficially seem lexically diverse if this study doesn't exclude these items from the count. By lessening the influence of function words, this study narrows down the lexical complexity principles to the real lexical wealth of the learners' lexical choices-nouns, verbs, adjectives, adverbs-as opposed to the omnipresence or repetition of grammatical particles. In the literature, this practice is suggested in order to acquire a cleaner signal in the use of vocabulary and is indispensable when calculating indicators such as the nominal ratio (the percentage of nouns vs. content words) or verb variation (the scope of verb usage), which reveal learners' stylistic and discourse tendencies.

Moreover, POS tagging is the foundation of a later dependency parsing, since often the accuracy and the outcomes of a parser are dependent on the proper POS tags. Thus, high-quality POS tagging also serves to directly aid lexical analysis along with syntactic analysis that will be used in conjunction with it.

Dependency Parsing for Complexity Analysis

Dependency contact techniques are typically used in analysis of syntactic complexity, providing information about the hierarchical structure of sentences, for example information about the grammatical functions of constituents (e.g., subjects, objects, modifiers) or about the linear distances between syntactically related units. Nevertheless, the same kind of analytical framework can help to enrich the studies of lexical complexity by elucidating the structural contexts within which lexical items will be used. Two main contributions emerge: First, the quantification of the construction of lexically complex elements in sentences, as the case of the chains of adjectival modifiers or the nesting of depends. Secondly, the calculation of some indices, such as the length of the dependencies, a syntactic variable although structurally it can reveal cognitive load associated with processing of the lexicon of the sentence.

One salient use of dependency parsing is related to the identification of sophisticated phrasal configurations. Consider nominal groups: In an advanced writer's hands a noun phrase like the significant long-term socio-economic implications of the policy may arise, in which several pre-modifiers and hyphenated compounds precede the head noun implications. A dependency analysis would show the presence of a number of adjectives and compound noun heads attached to implications. Recognition of such constructions is important, as they often contain advanced vocabulary (e.g., socio-economic) and display lexical richness through combinatorial means.

Traditional lexical indices, which focus on individual word frequencies, may miss one important feature of words, in that words may form a complex idea if they are complex, when parsed. Zhang and Zhang (2025) believe that when syntactic features are measured at a fine granularity, this study acquires more knowledge about the structure of language. In the research

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

on syntactic complexity, they support coupling automated dependency analysis and learner corpora for investigating finely graded aspects - specific dependent clause configurations, tree depth and so forth. This methodological refinement facilitates consistency and transparency between complexity studies.

Extending this inquiry to lexical complexity, dependency parsing makes it possible to report such measures of lexical complexity as the mean number of modifiers per noun and the frequency of multi-word expressions. Consequently, through the extraction of the syntactic conditions of the linguistic use of lexical items, scholars can better correlate the syntactic configuration to the richness and usage of the lexical item.

Another contribution to dependency parsing deals with the measurement of distance of dependence that has cognitive implications. Dependency distance refers to the linear distance between a head word and its dependent within a sentence, e.g., between a verb and its object. Greater distances generally present greater processing demands and could indicate more complex sentence structure. The relationship that the dependency distance has with lexical complexity can be seen by seeing that sentences full of lexical sophistication - dense in informational content - often entail elongated distances, as modifiers or subordinate clauses intercede between related words. For example, in the sentence the committee, wasted by the technical details, decided to postpone the decision, the participial clause intervenes between the subject head committee and the verb head decided, thus increasing the dependency length. Lei and Wen (2020) in a diachronic corpus study documented temporal decline of dependency distances in the written English and interpreted the result as evidence of a trend to more efficient, less complex syntax in some genres. In contrast, learner language tends to have short distances at low proficiency levels (in accordance with the simplicity of left-branching structures) and long distances at advanced proficiency levels (with the frequent inclusion of embedded clauses and parenthetical structures). While dependency distance is essentially a syntactical measure, it is secondary an indicator for lexical deployment: if a learner is able to hold a sophisticated phrase in working memory while filling a coherent sentence in terms of syntactical coherence, the mastery in both syntax and vocabulary is given.

Further, by using dependency parsing, it is possible to identify subordinate clauses and the lexical content they contain. By parsing sentences, researchers can see whether a clause is embedded, for example, a relative clause describing a noun or a clause functioning as a noun phrase. Advanced writing often combines complex information through subordinate clauses, which means that more words are introduced, including relative pronouns and different verbs and other function words. As a result, there is a way to measure subordinator clauses density for each sentence and the lexical items that are used in subordinate clauses. Learners that use more subordinate clauses tend to have more lexical diversity in connective and relative pronouns as well as more diversity of verb explained by clause-specific predicates. Blinova and Tarasov (2022), focusing on legal texts, presented a hybrid model of text complexity estimation using a combination of a linguistic feature extractor (larger set of features also based on dependencies between dependent features) and a neural network. Their model included both traditional lexical indications as well as syntax, thus reflecting the integrated approach espoused here. The high accuracy of the model shows that syntactic dependency information combined with lexical features provides a more powerful evaluation of textual complexity. In learner-"language analyses, this means that parsing can help us to understand why a text is considered complex, not just because it has difficult words in it, but because it contains difficult words embedded in embedded or extended syntactic structures, adding to overall complexity.

Specifically in regard to learner corpora, Zhang and Zhang (2025) offer a relevant example, although their main preoccupation is with syntactic complexity. Their research

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

combined automated dependency parsing with a self-built learner corpus, in view of extracting detailed metrics - including those on counts of complex nominals and frequencies of dependent clauses. They said this approach was key in improving granularity, accuracy, consistency, and transparency in measuring complexity as it offers an answer to perennial challenges in complexity-accuracy-fluency research. By analogy, applying dependency parsing in lexical complexity investigations should also be expected to lead to improvements in measurement - that is, by providing more granularity, as shown by more metric and contextual information, greater accuracy as a result of the correct attribution of word relations, and the avoidance of errant segmentation, consistency provided by a set of uniform parsing rules, and transparency in the shape of explicit articulation of the set of structural features taken into consideration.

Practically, in dependency parsing this study finds a set of metrics which are useful for the analysis of lexical complexity. These include: the average number of modifiers per noun phrase; the proportion of content words enclosed in multi-words expressions based on contiguous dependencies; the frequency of sophisticated lexical items (subordinate as opposed to main clauses); and the overall mean length of dependencies per sentence. Each of these metrics is then map-able to lexical complexity. For example, a higher modifier: noun ratio means that higher learners enhance each noun lexically (e.g., complex geopolitical issues not just issues). Likewise, a high mean dependency length implies insertion of descriptive or qualifying material between core sentence elements indicating a more roundabout lexicon and stylistic sophistication.

Integrated Approaches and Lexical Complexity in Proficiency

The complexity of language on many levels has led to the growth of multiple integrative or hybrid analytical frameworks. These models combine a wide range of feature sets, such as lexical, syntactic, coherence and discourse-level features, and therefore, they produce a more complete representation of language proficiency and text quality. Theoretical justification for such an approach is provided by Complex Dynamic Systems Theory (CDST), which argues that language development follows non-linear trajectories, which are shaped by the interaction between lexis, grammar, and other elements of language in time (Larsen- Freeman, 2017). Empirically, the proponents of this paradigm, on the one hand, caution against an isolated focus on independent elements of language, on the other hand they argue that correlational and interactive analyses of lexical complexity together with the other variables allow a greater insight.

Crossley's 2020 review is a good example of such an integrated perspective. She argues that lexical sophistication, syntactic complexity, and cohesion are each important to quality assessments of writing, and she goes on to claim that the three features are often interactive. For example, a text with a sophisticated vocabulary but lacking in cohesiveness may rate less well in quality since readers will encounter difficulties in comprehending the text. Conversely, a text of moderate vocabulary can even astonish by being coherent and well structured. Crossley points out that, in general, higher proficiency writing is more linguistically sophisticated, has a more complex syntax and is more cohesive; although the intervisibility of the relationship is evident across different contexts. These results highlight the importance of looking at multiple features at the same time in order to gain a complete understanding of proficiency.

Recent hybrid modelling research supports this conclusion. Bulte, Housen, and Pallotti (2024) call for clearer conceptual distinctions between complexity and related constructs such as difficulty, but note that empirical investigations will need to use multiple metrics to capture the ability of the learner. They present a theoretical review that calls researchers to put together comprehensive methodological toolkits when measuring complexity. An important implication

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

is that lexical complexity should be coupled with functional competence - not the possession of sophisticated vocabulary in general but the effective and contextually appropriate use within complex structures. Such a theoretical position is the driving force behind the investigations which consider whether learners who make use of advanced vocabulary also tend to construct complex sentence structures - even if these attempts are not always successful. Evidence that supports this association would support the presence of an underlying proficiency factor that would drive both of these phenomena at the same time.

From the computational point of view, integrated models using machine learning have shown that the concatenation of feature sets has made better proficiency predictions compared to single feature models. Blinova and Tarasov's (2022) aforementioned hybrid model used features from both the lexical and syntactic categories, which outperformed models that used only one of these categories when predicting the complexity of the legal text. Likewise, in evaluation of second language writing, researchers have attempted to predict human-scored values in a second language writing task by training regressions and classifiers with assorted assemblages of features. For example, Janebi Enayat (2025) studied the issue of narrative essays in the context of EFL in an Iranian EFL setting by using a comprehensive battery of computational indices (both lexical and syntactic). The analysis found that lexical diversity and sophistication were among the most powerful predictors of human rated quality of writing, even when indices of syntax were also included in the analysis. Nevertheless, however, the study also found that certain syntactic metrics (e.g., manually coded error-free T-unit ratios) played an additional predictive role over and above lexical metrics alone. These results underscore the dangers that excessive focus on rare vocabulary in scoring rubrics can skew perceptions of quality; that is, that people may reward advanced use of rare lexicon and thereby ignore larger qualities of essays that are negative. Consequently, advanced vocabulary has a significant impact on how proficient one is (reflected in the high predictive utility), but it must be assessed in combination with clarity and accuracy.

Based on these insights, our study precisely combines the lexical and structural analyses. A more advanced learner will show evidence of complexity through the elaborated lexicon that is embedded in complex syntax. Using one measurement for each dimension helps capture performance in a more meaningful way. For example, two learners may receive similar scores on a vocabulary test, but in written production one can use the vocabulary in rudimentary sentences and the other can use it in elaborate arguments. An integrated analysis would probably place the latter's writing on a higher level. This seems to coincide with the findings of O'Leary and Steinkrauss (2022) who have noted that lexical and syntactic complexity both blossoms simultaneously at a higher level of proficiency after compensatory trade-offs at the lower level stabilize.

Finally, an integrated approach fits in with pedagogical approaches, for example, the Common European Framework of Reference (CEFR) that describes proficiency in terms of learners' linguistic performance in multiple domains. For example, one of the descriptors of a B2 (upper-intermediate) is that the learner "can use a wide range of vocabulary for matters connected to his/her field and most general topics, though some hesitation and circumlocution remain." This specification is implicitly considering lexical range (lexical complexity) besides fluency and compensatory abilities which relate to discourse competence. An integrated research analysis endeavors to imitate this integral outlook. Indeed, Zhang and Lu (2025) used CEFR classification alongside complexity features and were able to determine that combining 24 features, for lexical, syntactic, and others, produced the best level of discrimination between proficiency levels. Their success implies that a learner profile of complexity in a variety of dimensions is characteristic of particular proficiency levels for example a B1 writer may have

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

moderate lexical diversity, predominate in high frequency word use and simple short sentences, while a C1 writer has high lexical diversity, frequent low frequency word use and extended and structurally complex sentences. This composite profile is exactly what our study aims to outline through combined part (of speech) and dependency-based analyses.

Research Gap

Although progress has been made in measuring lexical complexity, few studies have integrated part-of-speech tagging and dependency parsing within a single analytical framework. Most prior research isolates lexical or syntactic measures, overlooking how advanced vocabulary interacts with grammatical structure. We have not yet explored how useful dependency parsing can be to study lexical syntactic ties. Furthermore, little work relates these computational measures with human ratings of writing quality. Thus, these measures would not be applicable pedagogically. The study gives a solution to all the limitations highlighted within the text. An integrated, corpus-based approach to measuring this complexity is proposed in this study. This approach combines POS tagging, and dependency parsing in order to capture nuanced, context-sensitive aspects of lexical complexity in learner writing.

Research Methodology

To answer the research question, this is a corpus-driven study to make extensive use of automated language analysis tools. In this section, this study briefly describe the research design, sampling technique, and data used and also the instruments/measures used (which pertain to the computational tools used for POS tagging dependency parsing) and theoretical framework this study adopt. A different section later in the paper addressed the key ethical issues addressed in the proposal. The methodology was designed to guarantee reliability, validity and relevance of findings for research and practice in applied linguistics.

Research Design

This study used the quantitative-corpus analysis design, which consists of the descriptive one and the inferential one. The design is comparative in nature in the sense that it will compare the two sets of learner texts (intermediate vs. advanced proficiency) for their differences in lexical complexity. It is also correlational in nature, and it will explore the relationship between the complexity measures of these learner texts and their external proficiency scores.

Corpus compilation: This study built a specialized learner corpus of English essays. The corpus consists of 300 essays (approximately 150,000 words in total). Half of these essays (n = 150) were written by intermediate-level English learners (around B1–B2 level on the CEFR), and the other half (n = 150) by advanced-level learners (around C1 level). All essays were written on comparable prompts in an academic genre (specifically, argumentative/opinion essays on general topics such as education, environment, technology) to control for topic and genre effects as much as possible. Each essay is an exam or assignment output with a length between 500–700 words, ensuring relatively uniform text length across the corpus (which helps in comparing measures like type–token ratio that are sensitive to length). The intermediate group's essays were drawn from a second-year undergraduate writing class for non-native English majors, whereas the advanced group's essays came from a graduate-level academic writing course for international students. This design creates a clear proficiency contrast, which is suitable for examining how lexical complexity measures differ.

In-depth investigation: The research design features an integrated analysis approach using POS tagging and dependency parsing on the corpus. After collecting the corpus, the study processed all the texts through an NLP pipeline (as elaborated under Instruments) that first tokenized and POS tagged the texts followed by dependency parsing. The texts in the pipeline produced an enriched version where each token is assigned with a POS tag and a dependency

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

relation. From these annotated corpora, this study extracted a range of lexical complexity indices (explained below). Some indices are *purely lexical* (e.g., type–token ratio among content words, percentage of low-frequency vocabulary), others are *syntacto-lexical* (e.g., average number of adjectival modifiers per noun phrase), and others serve as control or descriptive variables (e.g., mean sentence length in words, which though syntactic can influence lexical opportunity).

The research design also includes statistical analysis to address the research questions. It is composed of comparisons of means (using independent samples *t*-tests or non-parametric equivalents as needed) between intermediate and advanced groups for each complexity measure. Additionally, the study have a multiple regression analysis where the dependent variable would be an external proficiency/writing quality score and the predictors would be our lexical complexity measures. This allows us to see which aspects of lexical complexity independently predict writing quality when controlling for others, aligning with the integrated approach highlighted in the literature (e.g., seeing whether diversity or sophistication has a stronger unique effect).

Validity and reliability considerations: This study aimed to improve the reliability of our findings with the use of automated and objective measures on a relatively large sample of learner writing. The same measurement is applied to every essay and is automatically counted or judged with no human rater involved. Because the prompts and lengths are very similar, it increases the internal validity of between-group comparisons, which means that differences between groups are more likely to be the result of proficiency as opposed to something else, like prompt difficulty or text length. To ensure the validity of the proficiency grouping, we cross-checked with pertinent test scores or teacher assessments wherever available. The results indicated that the intermediate group corresponded to an average IELTS writing score of approximately 5.5–6, and the advanced group 7–7.5. That is, there was a significant proficiency difference between the intermediate and advanced groups.

In summary, design of the study might be described as a comparative corpus analysis using tools from computational linguistics that is set into a quasi-experimental design, because proficiency was not randomly assigned; rather two groups representing proficiency were naturally observed. The design is cross-sectional rather than longitudinal - that is, it compares levels at one point in time - rather than longitudinal, which is consistent with the goal of being able to capture proficiency-linked differences at high granularity. Grounded in the premise that the combination of lexical and syntactic analytical techniques can lead to new insights as well, the design is implemented in a systematic and replicable way through the use of attested (known) instruments of NLP.

Sampling Techniques

The sampling strategy utilized was a purposive and criteria-based strategy. The target population consisted of adult English learners who are located in the context of academic literacy, with a special emphasis on two different subpopulations of adult English language learners: intermediate and advanced second language English language writers. We drew our sample from academic programs to ensure the writing tasks would be comparable (academic essays) and the learners would have some homogeneity in educational background (all were university students, though from various L1 backgrounds).

Selection of participants/texts: Within the available pool of essays from the two courses mentioned (undergraduate and graduate writing courses at a large university), we applied the following inclusion criteria:

- The essay had to be written under timed exam conditions or as a take-home assignment without external assistance (to reflect the learner's independent language ability).

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

- The essay had to be original (plagiarism-free, verified via plagiarism checking software).
- For the intermediate group, we included only essays written by students who had completed approximately 2 years of English study and were rated by their instructor as "mid-level" in writing proficiency. For the advanced group, we included essays by students in an advanced academic English course, typically those who had already achieved an IELTS 7 or equivalent in writing upon entry.
- We aimed for a balance of L1 backgrounds to avoid bias (though our focus is not on L1 transfer in this study, we did not want all advanced to be, say, Romance language speakers and all intermediate to be another group). Our final sample included L1 speakers of Chinese, Spanish, Arabic, and French in both groups, among others. No single L1 constituted more than ~20% of either group, ensuring diversity.

Sampling method: We used a stratified sampling method. The strata were the two proficiency levels. From each stratum (the pool of available essays at that level), we randomly selected 150 essays that met the inclusion criteria. There were about 200+ eligible essays in each pool, so our 150 is a large portion, but we did randomize which specific ones were chosen to avoid cherry-picking particularly good or bad essays. This random selection within levels enhances the representativeness of the sample for those populations. It is worth noting that because this is not a probability sample from all learners worldwide, we do not claim broad generalizability; rather, it is an *analytic sample* aimed at internal comparison and relationship discovery.

Sample characteristics: The intermediate group (n=150 essays) had writers with an average age of 20. They had diverse majors (some were English majors, others from sciences but taking English as a required course). The advanced group (n=150) had an average age of 24, many being graduate students in various fields (engineering, business, etc.) who were non-native English speakers. The advanced group thus had more content knowledge in specific domains; however, our essay prompts were general enough (e.g., "Do the benefits of technology outweigh the disadvantages in education?") that specialized knowledge was not required, minimizing any advantage in lexical complexity coming from domain-specific jargon. If anything, advanced students might know and use domain terms, but since prompts didn't specifically elicit those, this likely did not skew results.

Corpus preprocessing: After selecting the essays, we removed any identifying information (names, mentions of specific institutions) to anonymize the data. Texts were converted to plain text format. Minor corrections were made only for obvious OCR or typing artifacts in a few cases (since some exam essays were scanned; we re-typed them carefully to have clean digital text). We did **not** correct learners' language in any way: spelling errors, grammatical errors, etc., were left intact because they are part of authentic learner output. The POS tagger and parser we used can handle minor misspellings by backing off to contextual cues, but there is an assumption that gross misspellings are few; in our corpus, spelling was generally decent (perhaps due to many being typed assignments where students used spell-check). In cases of unrecognized words (some proper nouns or jargon), the tagger marked them as "X" (unknown), and we excluded those tokens from analysis of lexical metrics to avoid distortions.

In summary, our sampling strategy was consciously aimed at creating clear differentiations in the level of proficiency and consistency across tasks. By carefully balancing important variables and using random choices within our subject populations, we attempted to assemble a corpus of learner's texts which is both of high analytic quality and which represents the proficiency levels being studied. The relatively large sample size of 300 essays has enough statistical power to detect differences and correlations on the effect sizes reported in previous literature (e.g., Crossley & McNamara, 2017 reported correlations between lexical features and

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

quality to range from 0.4 to 0.6; the current sample is sufficiently powered to reliably detect such effects at $p < .05$).

Instruments

This research paper made use of a combination of computational programs and analytical measures, all of which are discussed here and their use and applicability to the research objectives.

1. NLP Pipeline (POS Tagger and Dependency Parser): This is a pipeline which combines these NLP components to produce the output: a dependency graph with labeling of the words (with word labels) and links between the words (with word links). Our analysis was based on an NLP pipeline which processed the texts. We made use of the Stanford CoreNLP toolkit (v4.2) through a Python interface, which is a tested language analysis toolkit. Part-of-speech tagging was done using the Stanford POS Tagger with the pre-trained model of engish-left3words-distsim, and parsing syntax dependencies was done using the Stanford Neural Network Dependency Parser. The chosen tools have been chosen because of its good accuracy in English, the POS tagger has an estimated accuracy of 97 0 -98 -99 -percent on the Penn Treebank WSJ corpus and the dependency parser is also scored with a high degree of labeling accuracy. Their use provided credible tags to every word with regard to POS and the grammatical framework of the sentence. We set the parser to produce Universal Dependencies (UD) output, which is a standard format that helped to specify the extraction rules of specific syntactic relations, like identify all adjectives attached to nouns.

2. Lexical Complexity Indices: Based on the tagged and parsed output, we computed a range of indices. Key indices included: - **Type–Token Ratio (TTR) for content words:** the number of unique content word types divided by the total number of content word tokens in the essay, expressed as a percentage. Content words were defined as all nouns, lexical verbs, adjectives, and adverbs (POS tags starting with NN, VB (excluding modal verbs), JJ, RB in Penn tagset). We excluded pronouns, prepositions, determiners, conjunctions, and modal verbs as function words. This measure reflects lexical diversity while mitigating the inflation from function word repetition. We also computed MTLD, which is known to be less sensitive to text length. We used an open-source implementation of MTLD that goes through the text and calculates the length of word strings required to reach a stable TTR value. MTLD was calculated for comparison with TTR to see if patterns held regardless of the diversity measure. - **Lexical Sophistication (Frequency-based):** We measured sophistication by looking at the proportion of words that are "advanced." Concretely, we used the **SUBTLEX-US frequency database** (Brysbaert & New, 2009) as a reference; it provides a frequency rank for English words. We defined "advanced" as those falling outside the top 5000 most frequent lemmas. We lemmatized each content word (using the WordNet lemmatizer) and checked its frequency rank. We computed: (a) percentage of content words in each essay that are in the lowest frequency quartile of the corpus (to capture use of rare words relative to this learner corpus's own distribution), and (b) percentage of words not in the top 5000 of SUBTLEX (to capture use of generally low-frequency words). Additionally, we incorporated an **Academic Word List (AWL)** measure: percentage of words from the AWL (Coxhead, 2000) present in the essay. This indicates use of academic vocabulary. - **Lexical Density:** Calculated as number of content words divided by total number of words in the essay. While simple, this needed tagging to compute properly. We expressed it as a percentage. This measures how information-dense the language is (high density means fewer function words per content word). - **POS-specific diversity:** We computed separate TTRs for nouns, verbs, adjectives, and adverbs. E.g., noun TTR = (unique noun lemmas / total noun tokens). This helps identify if a student's diversity comes mostly from one word class or is spread out. - **Mean Word Length:** Although not a

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

direct complexity measure, we included average word length in characters as a rough proxy for lexical sophistication (longer words tend to be less common). Prior work shows L2 writers at higher levels use slightly longer words on average. We computed this from raw text (excluding punctuation).

3. Dependency-Derived Measures: Using the parsed trees, we extracted: - **Mean Dependency Distance:** For each sentence, we calculated the average linear distance (in words) between heads and dependents for all dependency relations. Then we averaged these distances across all sentences in an essay to get a mean dependency length. A higher number suggests more complex sentence structuring (potentially embedding). - **Modifiers per Noun Phrase:** We identified all nominal heads (nouns that are roots of noun phrases) and counted their adjectival modifiers (amod relations in UD) and nominal compounds (compound relations). We then computed the average number of modifiers per noun in the text. For example, in "the two innovative approaches," "approaches" has 2 modifiers ("two" and "innovative"). We averaged such counts. A higher average might indicate descriptive richness (advanced writers often pile modifiers). - **Subordinate Clause Incidence:** We used dependency labels to count subordinate clauses, specifically the number of *adverbial clause modifiers* (advcl), *relative clauses* (acl:relcl), and *complement clauses* (ccomp and xcomp) per sentence. We then looked at lexical complexity within those clauses versus main clauses. However, for simplicity, our reported metric is "subordinate clauses per sentence" as a structural complexity indicator. - **Content Words in Dependents:** We analyzed the proportion of content words that were in dependent positions as opposed to head positions in the parse tree. The intuition: in a simpler sentence, the main structure might carry most content, whereas in a complex sentence, a lot of content is packed in subordinate dependents. This was exploratory.

4. Human Rating/Score: Although not an "instrument" in the same sense, an important variable was the **writing quality score** for each essay. For the exam essays, we obtained the official scores (on a 0–100 or 1–10 scale depending on context) assigned by instructors or examiners. We standardized these to a 0–10 scale for uniformity. These scores served as the criterion measure in correlational analysis. While subjective, they represent the kind of holistic proficiency judgment we're interested in predicting/explaining through lexical complexity. The inter-rater reliability where applicable (for the exam scripts double-marked) was above 0.85, indicating the scores are consistent.

5. Statistical Tools: For analysis, we used IBM SPSS (Version 27) for statistical calculations. Additionally, we utilized **R (Version 4.1)** with packages like *quanteda* for double-checking some corpus statistics and *ggplot2* for creating graphs (not the focus here, but to visualize distributions during analysis). Our statistical analysis included *t*-tests (via SPSS) and Pearson correlation and multiple regression (enter method) to examine predictive relationships. These are standard instruments for quantitative data interpretation in our design.

All instruments were calibrated and tested on a small pilot set of 10 essays beforehand to ensure they functioned as expected. The tagger and parser outputs were manually inspected for a couple of texts to ensure accuracy (they performed well, minor errors in tagging rarely affected overall indices because they were few). The measures we computed were also cross-validated: e.g., we compared our MTLD output on a text to known values from other tools to ensure correctness.

In summary, our instruments comprised a blend of NLP software for linguistic annotation and statistical tools for data analysis. This combination allowed us to systematically quantify the textual features of interest and relate them to proficiency, addressing our research question with rigor and replicability. The integrated use of these instruments is a strength of

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

our methodology, as it reduces human error and subjectivity, and enables analysis of complex interactions within the language data.

Theoretical Framework

The theoretical framework on which our study will be built combines the knowledge about the Complex Dynamic Systems Theory (CDST) and corpus linguistics, thus creating a hybrid approach to the analysis of the language complexity. This framework is a conceptualization of language proficiency as an emergent characteristic of various interacting subsystems; lexical, syntactic, morphological and others, who change over time in a non-linear manner. In this paradigm, lexical complexity does not occur as a singular characteristic but instead it evolves alongside other aspects of language. We combine this theoretical with an empirical approach that is corpus based and thus the theory, which is corpus based, matches the theory with the corpus data.

One of the latest models that guide our strategy is the hybrid model of complexity estimation suggested by Blinova and Tarasov (2022). Their study on legal texts, however, has lower principles that can be applied elsewhere. Their hybrid model combines the traditional elements of linguistics and the machine learning, i.e., neural networks, in forecasting the complexity of the texts. The success of that model highlights one of the central theoretical arguments: a combination of characteristics, that are borrowed at various levels of linguistics, will create the best estimation of the overall complexity. This observation in the current context makes the use of lexical and syntactic analyses jointly acceptable. Such progress, as the theoretical justification is that the competence of an advanced L2 writer is expressed through a conglomerate of actions, i.e., the use of infrequent lexical elements, construction of complicated sentences, and the existence of cohesion in a way that emphasis on a single area would under-represent the proficiency. Accordingly, our theoretical approach assumes that the lexical complexity must be studied as a component of a single system, which directly leads to our multi-faceted approach.

CDST provides the concepts of non-linearity and variability in development. As an example, when a learner is introduced to a new area (e.g., learning many domain-specific lexical items in a short amount of time), the lexical diversity can grow suddenly, whereas the syntactic complexity of the language can either stabilize or decrease in the short term (as the learner gets used to adding these new items to their lexicon). Gradually, however, both dimensions may be advanced. The model promotes the consideration of the fact that cross-sectional differences, e.g., those that are depicted between intermediate and advanced groups, are only snapshots of a fluent process. We discuss our findings as follows: the differences we have observed are probably due to where the learners are in some developmental path as influenced by instruction, exposure, and use, and not necessarily by some traits. Through this view we reduce the chances of overgeneralization; say we find that advanced learners use more adjectives, we can be reminded by CDST that this trend may be a stage of development and not necessarily the limit of all learners.

Another component of our framework is the Noticing Hypothesis and Input Enrichment from SLA theory, as it relates to lexical acquisition. Learners develop lexical complexity partly by noticing new vocabulary in input and incorporating it into output. Our focus on part-of-speech and dependency relations also connects to how learners chunk input; for example, by noticing collocations or syntactic frames associated with new words. A theoretical hybrid model by *Verspoor, Schmid & Xu (2017)* (in a CDST approach to writing development) suggests that *writers gradually enrich their lexicon and syntax and the synergy between these leads to more advanced writing*. We interpret our POS+dependency integration through this lens: a learner's ability to use a sophisticated word appropriately in a complex sentence signals

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

a higher level of development than using it in isolation or not using it at all. The theoretical framework thus predicts that measures capturing such appropriate usage (like our dependency-linked metrics) will better discriminate proficiency or predict quality than those that do not.

In terms of conceptualizing proficiency and writing quality, we lean on theoretical models of L2 proficiency such as Bachman's model (which includes organizational and pragmatic knowledge) and specifically the notion that *organizational knowledge* includes grammatical and textual knowledge – covering vocabulary and grammar. Our hybrid analysis touches on lexical and syntactic levels, thus keeping in line with the theoretical view in question. The inherent assumption is to have an efficient writer come up with a well-organized writing due to the advanced use of vocabulary and correct grammar. As a result, our theoretical stance is that the organization of the texts using lexical complexity, including its clarity, precision, and nuance, and syntactic dependency structures, including their mechanisms of linking ideas together, determine the way in which the text is organized. This rationale is the reason why our empirical research goes beyond a list of high-frequency words, as we are allowing their distribution into syntactic structures through a framework of dependency.

In short, this study comes up with a hybrid approach to second-language (L2) complexity that combines both components of a complex systems theory, with its focus on holism, subsystem integration, and interaction dynamics, with a corpus-based, computational measurement model. The most recent academic work, especially the efforts of Bulté and Housen, who differentiate between complexity and difficulty and Verspoor et al., who use CDST to analyze written discourse, proves this belief, as they are the first in the ongoing discussion to implement these metrics to position the area of study as consequential and essential. In practice, this can be translated into an examination of lexical complexity that lies within the context of syntactic packaging and communicative effectiveness in general. In this regard, our framework supports the fact that many indices were included and the effect of their overall impact on the writing quality was analyzed.

This new hybrid orientation does not stick to a specific disciplinary camp, to either a pure cognitive, pure social or pure structural approach, but instead borrows epistemic resources of computational linguistics as the tools of analysis, and the second language acquisition (SLA) theory as the interpretive framework of developmental paths and the competence of the learners. The structure goes on to provide answers to more specific questions than the question, how many high-frequency words does an advanced learner use? to more specific questions like, Under which contextual conditions do advanced learners make use of complex lexical items, and in what ways does this interdependence with their syntactic structure contribute to the efficacy or high-status of their writing? It also dictates result interpretation: e.g., in case of limited lexical diversity among advanced learners, CDST would alert one to either the possibility of developmental lag in the latter or of the possibility of the latter being offset by other linguistic properties, such as cohesion.

To sum up, our theoretical framework can be described as a perspective of complex dynamic systems of learner language complexities operationalized by a corpus-driven hybrid model. This approach recognizes that language proficiency is compound, and integrative analysis is required, hence making concrete in our application of multi-level methods of analysis. By placing our study within this context, we will not only attempt to add empirical findings, but also present our methodological paradigm that will exemplify how learner corpora may be studied in harmony with current theoretical ideas of SLA as an interactive process.

Research Ethics

The current study was conducted within the set of institutional and disciplinary ethical principles regarding the work on applied linguistics. Since the information contained written

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

reports in the coursework and examination scripts by the students, the informed consent of the participants was obtained, which guarantees voluntary participation to the study and no negative impact on their academic performance. In the case of the essays of former students unreachable, the Institutional Review Board (IRB; Protocol #2025-17) approved the usage of anonymized information with a minimal-risk exemption. All documents were coded using identifiers, personal references were excluded and results were only reported in aggregate form so as to maintain confidentiality. Data was stored on secure encrypted devices under passwords and all data processing done at local level without any transmission to other servers. The use of graded coursework and examination materials were acquired through institutional permission, and thus this ensured that academic integrity and intellectual property policies were abided. The research had low risk to the subjects since it focused on linguistic structure and not any personal content. All in all, the study met high ethical principles of ensuring privacy, data security, and respect and transparency to all participants.

Results and Findings

This part outlines the findings of the current research, based on the key elements of the research question: (a) differences in lexical complexity between intermediate and advanced learner corpora revealed in the number of metrics, (b) conclusions drawn in the context of part-of-speech tagging and dependency parsing to the patterns of the lexical items use, and (c) the correlation between these complexity measures and writing-quality scores. Takeovers and mergers 1 and 2 give brief displays of the salient quantitative results in APA format. Statistical operations were carried out at a traditional 0.05. Altogether, the findings show that there are strong differences in lexical proficiency levels in various aspects of lexical complexity, which accentuate the virtues of a combined analytical approach. The more sophisticated lexical repertoire was used by the more advanced learners and there is a higher likelihood of advanced learners placing the lexicon in more complex syntax structures. More so, a large percentage of the complexity indices showed high correlations with rubric-based essay scores suggesting that they incorporate consequential features of writing competence.

Lexical Complexity Profiles of Intermediate vs. Advanced Learners

First, descriptive statistics and comparative analysis of the key indices of lexical-complexity between the two proficiency groups are provided. Table 1 provides the group means (with standard deviations in parentheses) of each metric and the resultant t -tests.

Table1

Descriptive Statistics and Group Comparisons for Lexical Complexity Measures by Proficiency Level (N = 150 each group)

Measure	Intermediate (B1/B2) Mean (SD)	Advanced (C1) Mean (SD)	t (df=298)	p-value	Cohen's d
Content Word TTR (percent)	47.3 (8.5)	57.9 (9.1)	–10.45	< .001	1.19
MTLD (lexical diversity)	95.6 (15.4)	121.3 (20.2)	–12.34	< .001	1.42
Lexical Sophistication (% advanced words)	12.4 (5.0)	twenty 19.8 (6.3)	–11.19	< .001	1.29
Academic Word List (% of tokens)	4.8 (2.1)	7.5 (2.4)	–10.30	< .001	1.18
Lexical Density (% content words)	49.7 (4.4)	52.1 (4.0)	–5.04	< .001	0.58

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Measure	Intermediate (B1/B2) Mean (SD)	Advanced (C1) Mean (SD)	t (df=298)	p-value	Cohen's d
Mean Word Length (in letters)	4.55 (0.40)	4.97 (0.45)	–8.40	< .001	0.97
Noun Diversity (noun TTR)	45.2 (10.6)	53.1 (11.8)	–6.27	< .001	0.72
Verb Diversity (verb TTR)	38.5 (9.7)	46.0 (10.5)	–6.65	< .001	0.77
Adj/Adv Diversity (modifiers TTR)	27.4 (8.3)	36.6 (9.5)	–8.81	< .001	1.02

Note: *All differences are statistically significant at p < .001. Effect size (Cohen's d) interpretations: ~0.2 small, ~0.5 medium, ~0.8 large.*

Advanced learners are much more complex in their lexical features in all measures as shown in Table 1. In terms of lexical diversity, advanced essays have a mean content-word Type-Token Ratio of about 58 the intermediate essays have a mean value of 47, resulting in a large effect size (d = 1.19). The Measure of Textual Lexical Diversity (MTLD) also demonstrates significantly more richer vocabulary in advanced writing (MTLD ⁻ 121 versus 96). These results verify that more competent writers make use of expanded usage of lexical materials. The even greater difference is found in lexical sophistication: in an average case, advanced essays have about 19.8 percent of the total number of the low-frequency or advanced vocabulary tokens, while intermediate essays have approximately 12.4 percent of such tokens. This variance (d = 1.29) highlights the fact that not only more words are used among advanced learners, more uncommon or specialist words are also used. The Academic Word List coverage of advanced writing is around 7.5, much greater than the 4.8 of the intermediate writing, which is indicative of more academic vocabulary and formal vocabulary being incorporated by the students at advanced stage. The advanced group has a slightly higher lexical density (52.1 49.7 percentage). The effect size is not large (d 0.58), though it is statistically significant (p 0.001).

The two groups show densities of about 50% which is consistent with past studies which show that lexical density might not be an effective proficiency distinguishing factor in similar topics. That is, intermediate students create the equivalent amount of content words, although the quality and range of such content word are different. This trend supports the finding of Yang and He (2025) that the lexical density can reach a certain level (possibly because of the essay genre need; both texts must have certain function words to be coherent and grammatically correct) and not increase any further. There are higher word lengths in advanced texts (4.97 letters versus 4.55 letters) and this is also consistent with the preference of the advanced students to use longer, multisyllabic words - usually academic terms of Latin origin. It is a crass measure, but it supports the general trend of increased lexical sophistication. A part-of-speech disaggregation of diversity indicates that more advanced writers have more variability in nouns, verb and modifiers (adjectives/adverbs). All these differences are statistically significant (p -0.001). Adjective/adverb diversity is the most significantly growing by relative terms (36.6 versus 27.4, d 1.02). It means that advanced writers not only have more an inventory of nouns and verbs, but also a more extensive range of modifiers, which allows them to describe things more deeply and in greater detail (e.g., an intermediate writer can use the word very and the word big repeatedly, an advanced writer can use the words highly, significantly, substantial or considerable). This observation is consistent with the later ones on dependency-based modifiers.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Overall, these comparisons show that higher L2 writers win out over intermediate L2 writers in the lexical complexity in most measures. They use a richer and more scholarly vocabulary and the content-word densities of both groups are similar. This empirical study confirms the effectiveness of corpus categorization by proficiency and builds up on the previous studies of the development of lexical richness.

Insights from POS Tagging and Dependency Parsing

After the analysis of POS-tagged and dependency-parsed analyses, we consider the way learners apply their vocabulary into the context. The earlier diversity measures were informed by POS tagging since it isolated content words and allowed POS specific counts. In this case, we consider the tendencies in content/function word usage, certain POS distributions, and structural synthesis of vocabulary into dependency parsing. A notable frequency of POS tag findings is that the percentage use of adjectives and adverbs among the advanced learners is relatively high on frequency compared to the intermediate learners. Adjectives and adverbs make up 7.4 and 5.1 percent of tokens respectively in intermediate essays.

Compared to this, in advanced essays, there are 9.8 percent adjectives and 6.7 percent adverbs, which is statistically significant (p is less than.01 based on the chi-square of the number of tokens). This is in line with the observation that high achievement students give more detailed descriptions and nuance, as it is also evident in the greater diversity of modifiers. On the other hand, the noun/verb ratios are not significantly different (nouns 25 percent in both groups; verbs 15 percent in both), and even the percentage change in advanced essays is only 12 percent, which is not significant (p.05). To this end, both groups equalize the amount of space they use to represent entities and actions but more space on the entities and actions through modifiers is given by the advanced writers. This quantitative POS result supports qualitative impressions often observed in instructor feedback: middle-range writing can sound more basic or blunt and higher-order writing can sound more descriptive or scholarly and is partially explained by the use of adjectives and adverbs.

Dependency parsing provided several revealing metrics about structural usage of lexis. Table 2 presents key dependency-related measures for each group, along with significance tests.

Table 2

Dependency Parsing Metrics by Proficiency Level

Structural Metric	Intermediate Mean (SD)	Advanced Mean (SD)	t (df=298)	p-value	Cohen's d
Mean Dependency Distance (words)	2.41 (0.35)	2.89 (0.40)	–11.60	< .001	1.34
Modifiers per Noun (avg. # of adjectives/compounds per noun)	0.94 (0.21)	1.31 (0.28)	–13.19	< .001	1.52
Subordinate Clauses per Sentence	0.58 (0.30)	0.85 (0.33)	–7.47	< .001	0.86
Content Words in Dependent Clauses (% of content words)	22.5% (8.0)	30.4% (9.5)	–7.47	< .001	0.86

Note: Dependency distance is measured in number of words between heads and dependents on average. Modifiers per noun include adjectival and compound noun modifiers. Subordinate clauses count includes finite and non-finite adverbial and relative clauses.

From Table 2, advanced writers show a significantly larger **mean dependency distance** (2.89 vs 2.41 words). This indicates that in advanced writing, on average, words that are syntactically related (like subject-verb, verb-object, noun-modifier) are further apart in the

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

linear sentence order, often due to inserted phrases or clauses. In practical terms, advanced sentences tend to be longer and more complex, with qualifiers inserted between core elements. An example: Intermediate might write "Many people *believe* this." whereas an advanced writer might write "Many people, *perhaps influenced by recent events*, firmly believe this." The advanced sentence has greater dependency distances because, say, the subject "people" and verb "believe" are separated by an inserted clause. Our results quantitatively confirm that advanced texts have this characteristic of more spread-out dependency relations (the large effect size d = 1.34 underscores a substantial difference). This aligns with higher structural complexity and reflects how advanced writers manage to incorporate more information per sentence.

Modifiers per noun is markedly higher for advanced group (1.31 vs 0.94). In other words, intermediate learners use on average just under 1 modifier (often none or one) per noun, whereas advanced learners often use at least one and frequently two modifiers per noun phrase. For example, an intermediate writer might say "*students* face problems" vs an advanced writer "*undergraduate students* face *significant financial* problems" (here "students" has one modifier "undergraduate"; "problems" has one modifier "significant financial" which is actually two stacked modifiers). The advanced example yields a higher count. This corroborates our earlier finding of more adjectives/adverbs: dependency parsing shows those adjectives are indeed being used as noun modifiers more often. The large difference (d = 1.52) suggests this is a strong marker of advanced writing – richly expanded noun phrases with multiple pre-modifiers or compounds. This feature contributes to the denser informational content and a more academic style, since academic English often employs noun phrases with multiple modifiers (e.g., "a *comprehensive international educational reform* initiative").

The ratio of subordinate clauses constituting a sentence is significantly greater in advanced essays, with the mean of 0.85 as compared to intermediate texts 0.58. This shows that the average number of subordinate clauses in sentences used by intermediate writers are just above half, whereas in sentences by advanced writers there is an average of about 85 - percent clauses used in the same sentence, and more than one subordinate clause is used in a sentence on average. Subordinate clauses include relative clauses, adverbial clauses (e.g. clauses starting with because or although), and complement clauses. Highly developed writers use more relative-clause structures like There are problems, which teachers have to overcome, and more complicated structures like There are disadvantages, but it is generally believed that …. Intermediate writers, on the other hand, prefer simpler compound sentences or even dispense with the subordinate clauses. The difference between the advanced writers and the basic writers (effect size d acquired 0.86) suggests that the advanced writers use clauses in a greater manner, thus, adding complexity to the syntax and incorporating more information or arguments into the single sentences. Such a tendency can be considered as typical of writing competence.

Finally, the percentage of the number of content words placed in subordinate clauses was also determined to measure the degree of lexical weight carrying of subordinate clauses. In advanced essays, the content in subordinate clauses is about 30.4 per cent but in the intermediate ones, it is about 22.5 per cent. This implies that high-level writers incorporate a significant portion of the central information and descriptive detail in sub-constructions- a fact that is demonstrated by the phrase, perhaps influenced by recent events that have its content words that expound on people. In comparison, intermediate learners depend more on the use of main clauses to communicate meaning, and this factor could lead to more linear sentence structure (in general one main idea per sentence). The difference in content-word distribution is in line with the frequency of subordinate clauses with a reported correlation coefficient r0.79.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Yet, this measure, though innovative, highlights the syntactic complexity of the lexical one in advanced writing: advanced writers use discrete and complex lexical items in a preferential distribution among subordinate phrases and clauses to play up their sentences.

On the whole, the dependency-parsing analysis shows that not only do more advanced learners have a more extensive lexical repertoire, but also make use of that lexicon in the structurally more complicated situations: they make longer sentences and use more embedded information in them and elaborate noun phrases. The intermediate learners can use some complex structures but to a significantly lower degree. Such findings coincide with the arguments about the complexity-comprehension trade-off: simple syntax is more likely to be present in the language of intermediate writers, as they prefer to focus on content delivery, and more complex syntax and lexis are achieved by advanced writers at the same time.

Another shade is seen in the context of lexical density. In spite of the similarity in lexical density among the groups', more sophisticated texts have high prevalence of modifiers and subordinate clauses. This seeming consistency in density can be explained by the respective augmentation in the volume of the function words utilized by progressive writers the use of conjunctions and relative pronouns, which are to connect clauses and support articles in the name of the nouns. As a result, the content word ratio does not increase significantly even with an increase in content volume in the form of modifiers. This trend indicates a balancing process: more sophisticated writers increase the informational material and, at the same time, increase the number of the function words, which make grammatical cohesion possible, and avoid a significant change in the percentage of density. This kind of dynamics can continue to be unseen without the analysis of parsing, which demonstrates the explanatory nature of combined linguistic studies.

Relationship Between Complexity Measures and Writing Quality

Lastly, we examined how the various complexity measures relate to the **human-rated quality scores** of the essays. Recall each essay had a score (standardized 0–10) from instructors. We computed Pearson correlation coefficients between the measures and the scores across the combined dataset (N=300). Selected results:

- Content word TTR: r = +0. fifty59, p < .001.
- MTLD: r = +0.60, p < .001.
- % Advanced words: r = +0.55, p < .001.
- Modifiers per noun: r = +0.58, p < .001.
- Mean dependency distance: r = +0.52, p < .001.
- Subordinate clauses per sentence: r = +0.45, p < .001.
- Lexical density: r = +0.18, p < .01 (significant but small).
- Mean word length: r = +0.49, p < .001.

All the main lexical diversity and sophistication indices had moderately high positive correlations with the quality of the essay. This indicates that the compositions with more and more sophisticated vocabulary stood higher chances of being rated higher. As an example, the vocabulary diversity (MTLD) showed a correlation of $r \approx 0.60$, which means that lexical variety is one of the most effective individual predictors of quality among the measures. Lexical sophistication, operationalized as high frequency of use of words, also exhibited a strong relationship ($r = 0.55$), suggesting that assessors reward the use of high-level lexical items, intentionally or not, since their use would make a piece of writing sound more fluent or adult. These results are congruent with the study by Lu and Hu (2022) and Crossley and McNamara (2016), who also found the predictive relations between lexical sophistication indices and quality ratings.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Importantly, the integrative steps, i.e., number of modifiers per noun, dependency distance also showed significant correlations (r = 0.58 and r = 0.52, correspondingly). This trend shows that the effect must not be attributed to lexical concentration and to the usage of advanced language in structurally complex phrasal constructions only. Compositions with denser modified noun phrases (such as severe economic and social issues and economic issues) and with a longer dependency structure, which is more indicative of a greater degree of embeddedness of information, were more likely to be rated higher. These findings could indicate the tendency of raters to rate as more developed texts that incorporate complex lexicon in complex and coherent grammatical systems. The frequency and quality correlation of the subordinate clauses were also statistically significant although slightly lower (r = 0.45), indicating that the use of the subordinate clauses leads to the higher quality, but too many convoluted sentences may lead to diminishing returns, even though complex sentences will potentially be confusing the raters.

Lexical density, on the contrary, turned out to have weak correlation with quality (r = 0.18), which highlights the inapplicability of the simple information packing. This is correlated to previous findings that indicated that both high and middle-level writers had similar lexical density. Raters seem to be more concerned with the expressiveness of ideas, clarity and subtlety, than with raw information. The density of an intermediate writer may be very high, by the use of telegraphic, functionless prose; but this terse style often does not contribute to the clarity, and therefore to the total mark. Thus, lexical density is not a good measure of the quality of writing in our sample.

In order to test the joint effect of the leading variables, the multiple regression model was formulated where the essay score served as a dependent variable and a set of predictors included in the model, content-type-token ratio (TTR), proportion of advanced words, the number of modifiers per noun, and the mean dependency distance. The criteria used in the selection of these variables was their representativeness and non-collinearity as indicated by the variance inflation factors of less than 2. The model produced a large F value, F(4, 295) = 89.5, p =.001 and accounted the 55% of the variance (R 2 =.55). Each of the individual predictors was significant (p <.01), and modifiers per noun had the highest standardized beta coefficient (0.30), then content TTR (0.27), proportion of advanced words (0.22), and average dependency distance (0.18). These findings show that both variables have a unique predictive value to quality (syntactic lexical integration has a slightly more significant effect compared to lexical diversity or the use of a rare word).

Simply put, an essay with high quality in our dataset is characterized as one with a large lexical repertoire, a discriminating use of words of high level, and their ability to integrate well-formed phrasal and clausal constructions. On the other hand, poorer essays are characterized by more repetitive and basic words and simpler sentence structure, so the prose of the poor-quality essay will look crude or immature in the eyes of the raters. Such results are promising in terms of education (since they define areas of pedagogical intervention, i.e., to teach students not only new words, but the strategies of their use in the complex noun phrases and sentences). Besides, it is clear that the presented usefulness of a multidimensional analysis methodology would be blurred by the use of only one measure, which would not be able to reflect the entire range of the factors that make language excellence.

To sum it up, the current experiment provides a high-quality empirical data that lexical complexity, especially when combined with the syntactic richness, is closely linked with the increased writing competence. We have statistically shown that more advanced L2 English writers are more lexically diverse and sophisticated than intermediate ones and that they actively invest the lexical resources in more elaborate syntactic structures. The integrative

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

analysis based on part-of-speech tagging and dependency parsing was also helpful to demonstrate these patterns, which proves the idea that the joint lexical-syntactic approach is more informative than the analysis based on lexicon or syntax on its own. The findings will be contextualized by placing them in a larger theoretical framework in the Discussion section that follows and practical implications delineated and methodological limitations and future research opportunities recognized.

Discussion

The findings of this study offer several important insights into lexical complexity in learner corpora and demonstrate the value of a corpus-driven approach that utilizes POS tagging and dependency parsing. In this section, we discuss the implications of these results, connect them back to the literature, and explore what they mean for theory, pedagogy, and further research. We also consider the limitations of our study and potential avenues for future work.

Lexical Complexity and Proficiency

The outcome of our findings clearly shows that lexical complexity is a sound indicator of L2 writing competence. The more developed vocabulary was used by advanced learners as Type-Token Ratio (TTR) and Multilevel D (MTLD) scores were higher, and more complex lexical decisions were made, which were characterized by a higher percentage of low-frequency and academic words, in comparison to intermediate learners. This trend is concomitant with the intuitive aspect that the more learners advance, the larger the lexicon and more detailed word knowledge they possess which in turn, they can effectively utilize in their writing. The current research builds on this research question by measuring the extent of these differences in a controlled environment (the same genre and comparable prompts). Both the effect size of diversity and sophistication were found to be large indicating that the level of lexical difference between the two languages is not subtle. These results support the claims in second-language acquisition that the development of vocabulary, both breadth and depth, is a key factor in general proficiency (Nation, 2017) and align with the finding by Crossley et al. (2019) that lexical sophistication proves to be the most significant predictor of human ratings of writing quality, which are supported by our correlation and regression analyses.

Here, one interesting subtlety was noted: there was no significant increment in lexical density with proficiency. The observation that the content in the ratio of intermediate to more advanced grades did not have significant differences is an indication that the addition of more content is not the main strategy underpinning the improvement of advanced writers. Progress seems to be made by adding richness to how the content is delivered, by modifying it, being more specific and using other elaborative devices instead of crowding more words which represent content per sentence. This fact agrees with the research of Yang and He (2025) which revealed that L2 and native texts are similar in terms of density, and it brings up a significant pedagogical fact, that simply asking students to add more facts, which may increase density, will not result in better writing, but asking them to elaborate and refine will result in more modifier usage and clause embedding, without significant changes in density.

POS Tagging Insights – Use of Modifiers

Part-of-speech tagging revealed that high-level learners employ a great deal more adjectives and adverbs in frequency and variety than do intermediate learners. Such variance implies an evolutionary process in the ability or desire to change and explain. Writing in L2 in early stages exhibits paratactic utterances (e.g., The car is fast. It is gas consuming (it has to be), and advanced writing leans towards hypotactic, elaborative structures (The car, which is designed to be high-performing, burns a lot of fuel). This qualitative change is supported by our empirical data. It reminds Hinkel (2003, 2019) who mentioned that academic writing

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

written by experts is rich in nominal modifiers and that, at the initial stage, nominal modifiers are underused by non-native speakers (NNS). By the C1 level, learners seem to be even as they might be a result of academic training that points at the necessity of qualifying statements (with the help of hedges, approximators, many of them being adverbs) and of providing detail with the use of adjectives. This observation is connected with tradeoffs in complexity: the use of more modifiers can make a sentence lengthier and more syntactically complex (which is reflected in our dependency measures). Complexity appears to be something that advanced learners appear to be happy to contend with to get the precision right, but to strategic learners, it may be better to keep complex sentences simple and leave the excessive modifiers out. This avoidance may represent either strategic or linguistic constraint; an intermediate learner may not have adequate synonyms or descriptive words or may be unsure how to fit them grammatically, so he/she plays on the safe side.

Dependency Parsing Insights – Structural Complexity

The dependency-based analysis offered a further insight into the analysis to show how the better learners organize their lexicon in sentences. One of the main findings is that advanced writing has greater dependency distances and more subordinate clauses, meaning that ideas are more entrenched as opposed to being represented in sequence of main clauses. Cognitively, the pattern indicates that the more advanced the learner the more processing capacity or comfort they have regarding working with more complex structures; they are able to sustain coherence in a longer stretch of words connecting related ideas and follow through several sub-propositions within one sentence. In intermediate learners, long dependencies or deep nesting in sentences can place too much processing or production demand on them, limited due to the working-memory limitations in interpreting or syntactic generation of sentences (e.g., the Dependency Locality Theory). They are therefore more likely to default to shorter dependencies, which is, positioning subjects and predicates close together, not interrupting clauses and reduced subordinate clauses and prefer multiple simple sentences or coordinate structures.

According to our data, the ability to cross this threshold, i.e., make uninhibited use of the subordinate clauses, is one of the stages of increased proficiency. This is also consistent with studies by Norris and Ortega (2009), who discovered that use of subordinate clause helps differentiate between more competent academic writing, a process that occurs, ironically, to vanish in very advanced writing as scholars seek out phrasal complexity (nominalization) to our learners, who have not yet attained this advanced stage. The considerable degree of dependency measurements and quality scores also indicates that raters appreciate syntactic complexity to some extent which is meaningful. When properly applied, more complex sentences allow the ability to express relationships between the ideas (cause- effect, concession, condition etc.) more finely in a single cohesive discourse unit, which may make an argument more convincing or more advanced. Raters can unconsciously associate such complexity with competence and, thus, give them more points. However, there is a serious need to take precaution and many researchers advise that complexity just to be complex is not the goal; the highly twisting sentences may hinder the understanding. Here our advanced learners seem to bring about a balance in that subordination is applied well but does not blur out meaning on an average basis.

The intermediate learner may also make more run-on or spaghetti sentences in an attempt to overuse subordination when they lack the necessary skill to use it, thus, their frequency of using subordinate clauses is less- maybe they have been taught to avoid run-ons by concentrating on just one idea in a sentence so they can avoid errors. These differences in pedagogical focus are visible: in lower levels of language instruction sentence-level clarity

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

(which may favor simpler constructions) is encouraged, in higher levels more sophisticated association of ideas (Uccelli et al., 2013 of academic language development).

Integrated Nature of Lexical and Syntactic Complexity

One of the brightest results of our research is the fact that the idea of lexical and syntactic complexity co-evolving, and their combination characterize the advanced writing, is supported empirically. The advanced learners would use the subordinate clauses to add more advanced words, such as the descriptive clauses with high-level words. Their high content words in dependent clauses statistic is an indication that the contents are embedded in complex structures. Such observation proves an integrated point of view: vocabulary is not an independent inventory but is a part of building complex messages. It concurs with theoretical hybrid models (e.g., Blinova and Tarasov, 2022) that suggest making a combination of several language features makes a more detailed description of complexity possible. No single measure prevailed in our regression analysis; rather a combination of diversity, sophistication, and structural integration showed the greatest consistency in predicting quality, which indicates that proficiency is a multifaceted concept. According to the complexity theory (CDST), these results show that the learner has more interconnections in his or her language system, with lexical complexity allowing, and being allowed by syntactic complexity in a positive feedback loop (Verspoor et al., 2017).

At intermediate level, subsystems are seen to be loosely related, a learner may be able to memorize more advanced words and have no proficiency in using them in a complex sentence or the opposite. At the advanced level, these issues are combined: the learner has the ability to use the higher-order words in the appropriately complicated sentences. This integration is exactly reflected in our data; as an example, the correlation between the percentage of advanced words and advanced modifiers in a noun across the essays was $r^{-}0.50$, which means that the higher the lexical subtly, the higher the noun modifiers in the essays, or the other way around.

Implications for Teaching

Practical pedagogical implications can be drawn on the findings. To begin with, the high positive correlation between the measures of lexical complexity and the quality of writing indicates that vocabulary development (in terms of breadth and sophistication) must remain a priority of the educators and curriculum developers even in high-quality levels of writing. In writing education, in addition to fixing grammar and the composition structure, focusing on the so-called lexical enrichment, i.e., encouraging students to use accurate and diverse word selection, may lead to quantifiable quality improvement, as rated by such a system. Nonetheless, our research indicates another need: the contextual use of words: simple knowledge of an advanced synonym is not enough but rather, learners need to learn how to put it into a sentence in the correct place and how it may be integrated in a sentence with modifiers or complex constructions. This includes teaching of collocations and repeated academic patterns (e.g., It is widely recognized that... where both structure and lexical sophistication are co-occurring).

Second, the advanced students used more subordinate clauses and modifiers which means that middle levels learners would receive selective exposure to more complex sentences. The training plans may involve exercises involving sentence combinations with subordinators or sentence extensions with subordinate clauses and appositions so that the students would be able to learn embedding strategies. By the time learners are at high proficiency levels, as our findings indicate that they must be proficient in such complexity: scaffolded instructions can assist in this process.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Lastly, one should exercise care to ensure that complexity will not be prized at the cost of being clear. Since the level of lexical density did not vary significantly with proficiency in writing, more sophisticated writers gain the opportunity to use complexity without impeding readability. This teaching should thus be focused on the idea that complexity is a means towards a more articulate and subtle expression rather than it being an end to itself (Ortega, 2012). students ought to be instructed to keep sentence length where it is appropriate, and use complexity where it is discourse appropriate and where it is necessary to simplify the structure to achieve clarity.

Limitations and Conclusion

The corpus of the study was short academic essays. The results might not be applicable to other forms (i.e., narrative, technical writing) or to the spoken language, which might exhibit alternative lexical-syntactic patterns. Students were classified into general (B1/B2) and high-performing (C1). Even finer variations within these levels (low vs. high C1) may provide some more subtle developmental variations. Instructors have to mark with authenticity, but the content and coherence might be in addition to the linguistic complexity of the essays marked by them. Associations with writing quality are therefore not causal relationships with writing quality. There may be small errors introduced in automated POS tagging and dependency parsing particularly when working with learner text. Although the state-of-the-art tools reduce this problem, there may be instances of mis-tag or mis-parses which may have a minor impact on the results. The research did not look at the interplay between grammatical errors and complexity. The presence of corpora with errors may indicate the simplification that learners make in order to make fewer errors- a possible trade off in L2 production.

This paper has evaluated lexical and syntactic complexity of 300 English learner essays through POS tagging and dependency parsing. Findings indicate that writing is lexically rich and structurally more complicated as learners advance. More developed vocabulary, more modifiers, and more structured in embedded forms are used by advanced learners, whereas intermediate learners use a simple and linear sentence. These measures of complexity have a strong positive relationship with the quality of the human-rated essays, indicating that more valuable lexical and syntactic resources provide better writing by making it clearer and more convincing.

The multi-level, corpus-based approach was particularly helpful, showing the subtlety that would not have been detected with the single measure. The results affirm that competence is not so much about knowing numerous words and rules but rather integrating them to communicate information in an effective manner. Pedagogically the research recommends emphasizing on the development of lexical diversity and how to teach students to develop sentences in a manner that is meaningful- yet clear. In general, the study proves that higher learners do not only know more words but also make better use of them, which confirms holistic approaches to language acquisition and provides valuable clues to further studies, instruction, and evaluation.

References

Blinova, O., & Tarasov, N. (2022). A hybrid model of complexity estimation: Evidence from Russian legal texts. *Frontiers in Artificial Intelligence, 5*, Article 1008530. https://doi.org/10.3389/frai.2022.1008530

Bulté, B., Housen, A., & Pallotti, G. (2024). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning, 75*(2), 533–574. https://doi.org/10.1111/lang.12669

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing, 44*, 51–62. https://doi.org/10.1016/j.jslw.2019.03.004

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

Heng, R., Pu, L., & Liu, X. (2023). The effects of genre on the lexical richness of argumentative and expository writing by Chinese EFL learners. *Frontiers in Psychology, 13*, Article 1082228. https://doi.org/10.3389/fpsyg.2022.1082228

Holmberg Sjöling, C. (2025). The effect of lexical complexity on grading of Swedish EFL learners' texts during high-stakes exams. *International Journal of Learner Corpus Research, 11*(2), 245–275. https://doi.org/10.1075/ijlcr.23038.hol

Janebi Enayat, M. (2025). Computationally derived linguistic features of L2 narrative essays and their relations to human-judged writing quality. *Language Testing in Asia, 15*, Article 35. https://doi.org/10.1186/s40468-025-00374-9

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*(1), 120–141. https://doi.org/10.1111/modl.12447

Kyle, K., Crossley, S. A., & Berger, C. (2018). The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Lu, X., & Hu, R. (2022). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods, 54*(3), 1444–1460. https://doi.org/10.3758/s13428-021-01675-6

North, K., & Zampieri, M. (2023). Features of lexical complexity: Insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence, 6*, Article 1236963. https://doi.org/10.3389/frai.2023.1236963

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

Verspoor, M. H., Schmid, M. S., & Xu, X. (2017). A dynamic usage-based perspective on L2 writing. In G. Fogal & M. Verspoor (Eds.), *Complex Dynamic Systems Theory and L2 Writing Development* (pp. 227–241). John Benjamins.

Xu, Z., & Casal, J. E. (2023). Navigating complexity in plain English: A longitudinal analysis of syntactic and lexical complexity development in L2 legal writing. *Journal of Second Language Writing, 63*, 100947.

Yang, Y., & He, X. (2025). Lexical richness in Chinese university students' EFL writing: A corpus-based comparison. *Humanities and Social Sciences Communications, 12*, Article 1199. https://doi.org/10.1057/s41599-025-05560-x

Yoon, H., & Polio, C. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and task effect. *Journal of Second Language Writing, 37*, 1–12.