Erzincan Üniversitesi Eğitim Fakültesi Dergisi

Erzincan University Journal of Education Faculty

2024 Cilt 26 Sayı 4 (542-553) https://doi.org/10.17556/erziefd.1483282

Araştırma Makalesi / Research Article

An Evaluation of Syntactic Complexity, Lexical Diversity and Text Length as Predictors of Foreign Language Writing Scores

Sözdizimsel Karmaşıklık, Sözcük Çeşitliliği ve Metin Uzunluğunun Yabancı Dil Yazma Puanlarının Belirleyicisi Olarak Değerlendirilmesi

Zafer Susoy[1] **Gül Durmuşoğlu Köse**[2]

[1] Asst. Prof. Dr., Tokat Gaziosmanpaşa University, Faculty of Education, Tokat, Türkiye
[2] Prof. Dr., Anadolu University, Faculty of Education, Eskişehir, Türkiye

Makale Bilgileri

Geliş Tarihi (Received Date)

13.05.2024

Kabul Tarihi (Accepted Date)

15.11.2024

***Sorumlu Yazar**

Zafer Susoy

İngiliz Dili Eğitimi Anabilim Dalı, Tokat Gaziosmanpaşa Üniversitesi, Tokat, 60250, Türkiye

zafersusoy@gmail.com

Abstract: The main premise of this study is to investigate to what extent syntactic complexity (SC), lexical diversity (LD), and text length (TL) correlate to foreign language (FL) writing quality scores assigned by human judges for the English essays of 204 pre-service teachers of English of two different curricular levels (first and fourth-year students). The study adopts a sequential-explanatory mixed-method research design. To that end, eight instructors rating student papers for 16 years on average were interviewed. The statistical analyses reveal that the 4th-year students outperformed the 1st-year students in TL, writing scores, and five indices of SC and LD. Subsequent regression analyses explained the variance in overall writing scores. The qualitative results showed variability in the instructors' ability to detect and prioritize these linguistic features, showing that while some instructors had a nuanced understanding of SC and LD, others emphasized overall content and organization more than linguistic complexity. The role of syntactic complexity, linguistic diversity, and text length as predictors of foreign language writing quality revealed that while human raters recognize these linguistic features to varying extents, their evaluation can be enhanced through standardized assessment practices and the integration of automated tools.

Keywords: Composition, proficiency, writing assessment, English as a second foreign language

Öz: Bu çalışmanın ana amacı, İngilizce deneme yazılarının değerlendirilmesi için insan hakemler tarafından verilen yabancı dil yazma kalite puanlarını ne ölçüde cümle yapısı karmaşıklığı (CYK), kelime çeşitliliği (KC) ve metin uzunluğu (MU) öngörebileceğini incelemektir. Çalışma, ardışık-açıklayıcı karma yöntem araştırma tasarımını benimser. Bu amaçla, ortalama 16 yıl öğrenci yazılarını değerlendiren 8 eğitmenle görüşmeler yapılmıştır. İstatistiksel analizler, 4. sınıf öğrencilerinin MU, yazma puanları ve CYK ve KC'nin toplam 5 göstergesinde 1. sınıf öğrencilerinden daha başarılı olduğunu ortaya koymuştur. Ardından yapılan regresyon analizleri, genel yazma puanlarının varyansını açıklamıştır. Nitel sonuçlar, eğitmenlerin genel puanlama prosedürüne ve CYK ve KC'yi puanlama sürecinde ne kadar iyi kavrayıp dikkate aldıklarına ilişkin içgörüler sağlamıştır.

Anahtar Kelimeler: Kompozisyon, yeterlilik, yazma değerlendirmesi, ikinci yabancı dil olarak İngilizce

Susoy, Z. & Durmuşoğlu Köse, G. (2024). An evaluation of syntactic complexity, lexical diversity and text length as predictors of foreign language writing scores. *Erzincan University Journal of Education Faculty*, *26*(4), 542-553. https://doi.org/10.17556/erziefd.1483282

Introduction

Writing instruction necessitates a process fundamental to a student's academic career. Students who struggle to express themselves clearly in writing may underachieve in class and possibly fail to graduate. Most of these risks come from high-stakes exams requiring advanced first-language (L1) writing abilities (Jenkins et al., 2004). Writing proficiency has a significant impact on academic success in L1 primary and higher education, as well as on future professional endeavors (Geiser & Studley, 2001). These advanced L1 writing abilities have been linked to sophisticated linguistic traits and linguistic elaboration (McNamara et al., 2010). Writing in a highly qualified manner in foreign languages (FL) has also been found to contain linguistic traits connected to more complex languages (McNamara et al., 2009). In the vast majority of prior research, the sophistication of language employed in written FL works, which influences writing quality evaluations, was primarily connected with syntactic complexity (SC) and lexical diversity (LD) (Crossley and McNamara, 2010, 2011; Ellis & Yuan, 2004; Lu, 2011; Mazgutova & Kormos, 2015). However, the notion of 'complexity' is a complex notion itself, in which complexity and diversity mostly overlap. According to Bulte and Housen (2012), L2 complexity has been handled from two basic perspectives: global complexity and local complexity. The first refers to the general L2 system of the learners and its dynamic nature, while the second refers to the particular items and structures. Following this distinction, we use the "global" view to define complexity and variety in our study:

> "*The learner's L2 system or "repertoire" that is, the quantity, variety, and richness of various structures and items the learner knows or employs—is referred to as global or system complexity. Examples of this include whether the learner is proficient in a small or large range of words or grammatical structures, whether he controls all or only a portion of the L2 sound system, and so on"* (Bulte & Housen 2012, *p. 25).*

According to Ortega (2012), up to now, the greatest number of researchers has concentrated on at least three key objectives while looking into complex issues in L2: "a) defining and measuring proficiency, b) describing and comparing performance, c) understanding how development proceeds concerning different factors such as age, initial competence, aptitude for language learning, and input quantity and quality" (p. 128). The threefold premise of the current investigation is similar. The learner's L2 system identifies correlations and determines how much a syntactic complexity and lexical variety—the quantity, variety, and richness of learner-knows structures and items the student uses, "repertoire"—is called global or system complexity (Hmelo et al., 2000). Examples can range from whether or not someone knows many or few words and grammatical structures in their second language while also taking into account how much of the actual sound system they have mastered. Global complexity or system complexity is when someone talks about the second language acquisition system or the "repertoire" of learners. Repertoire refers to the quantity, diversity, and affluence of different structures and items available in a student's knowledge base or active production process. For instance, examples are, if a learner can range through many or few words and/or grammatical structures, if either all or part of the L2 sound system is under his control or not. The student's language system in the foreign language or "repertoire"—meaning the amount, diversity, and number of distinct structures and more specific items that a learner knows in a given language—is termed as general or systemic complexity. For example, one learner may be proficient with only a few words or structures while another may know many more but still not all of them as well as he would like in speaking; still others may come close to mastering some features though always falling short somewhere else. The student's L2 system is known as system complexity at the global level or it is his/her or her global or systemic complexity. For instance, according to a paper by Muter et al. (2004), whether a learner is good at a small or big range of words or grammatical structures can be one of the examples. A different illustration is if he controls the overall sound system of L2. Global or systemic complexity indicates the extent, diversity, or richness of structures or items that the learner knows or uses in a second language. This is evident when the learner knows few or many words and grammar rules, controls some or all the sounds in L2, etc. To account for this diversity in writing quality among FLs, it examines how these factors relate to human ratings for FL writing quality.

The present study's second objective is to appraise the proposed connection between syntactic complexity, lexical diversity, and writing quality in second language learner writings from both developmental-based and proficiency perspectives in a more human-like manner that would appeal more to the general audience of researchers and educators who are working together with us on this project. (Crossley & McNamara, 2014; Ortega, 2012; 2015). The current study's third premise is to explore the perceptions of instructors related to SC and LD who have been scoring undergraduates' academic writing in an English Language Teaching Department. Thus, we aim to see the extent to which these instructors are aware of SC and LD in their scoring procedures. The findings and recommendations made by earlier studies are inconsistent and have a variety of flaws, including a small amount of data, learners' identical proficiency profiles, and insufficient sampling (Ortega, 2003). However, the present research suggests that using a reliable text-processing tool to incorporate several metrics across differing proficiency levels in one extensive dataset might offer a clearer view concerning how syntactic complexities are related to lexical diversities in the field of FL writing. In addition, our participating students are EFL pre-service instructors who are expected to instruct English language and FL writing at various levels, in contrast to the research that has been examined.

Literature Review

Measuring Syntactic Complexity and Lexical Diversity: Methods and Problems

Numerous measures of SC and LD have been proposed in the literature. For decades, there have been research initiatives to identify and validate a trustworthy measure of these constructs (Ortega, 2003; Wolfe-Quintero et al., 1998). The majority of this research has concentrated on identifying the measure(s) that might be objectively used to gauge writing proficiency, tracking SC and LD components in writing. The amount of data, the operationalization of the language tasks and genres in the data collection processes, as well as the variability and consistency of the complexity measures, lead to discrepancies in the findings of these studies (Lu, 2010; 2011; Ortega, 2003; Wolfe-Quintero et al., 1998). The inability to pool the findings of earlier studies is hampered by not just the inconsistent metrics utilized but also their scarcity and the small amount of available data. For instance, only four of the twenty-five cross-sectional studies evaluated in Ortega's (2003) thorough analysis of the development of syntactic complexity in writing in a foreign or second language that used four to five different metrics. Only three metrics were used in the other twenty-one investigations. The mean number of words in each written sample is 234 with a standard deviation of 110, while the average number of written data obtained in these investigations was fewer than 100. In subsequent work, the same issues persisted. For instance, in one study only clauses per-T Unit measures were used to syntactically examine 300 learner emails (Stockwell & Harrington, 2003). In a different study, Beers and Nagy (2009) employed mean clause length in addition to the T-unit ratio to assess 41 essays in two different genres, Ellis and Yuan (2004) used only clauses per T-unit measure to analyze 52 narratives. Text length as a measure of syntactic complexity, however, poses serious problems of reliability. Although text length was often associated with overall writing quality scores assigned by human judges (Guo et al., 2013), some other studies showed that text length does not necessarily increase along with syntactic complexity indices (Becker, 2010; Stockwell, 2005)

Also, problems arise when trying to measure lexical diversity. The first method applied in the past in the measurement of lexical diversity is several different words (NDW). The major challenge for NDW is that it relies too much on text length. "Most probably the number of different words in a language sample will depend on the total number of words in total" as stated by Malvern et al. (2004). This is the fundamental problem facing lexical (vocabulary) diversity measurements" (p. 16). Another one of the most widely applied lexical diversity measures is the type-token ratio (TTR). While type counts the variety of words in the text, token counts the entire number of words in the text. Thus, it has been recommended to employ a type/token ratio to enhance the reliability of NDW. It is more precise to calculate a ratio instead of just counting unique words, but TRR suffers from the same problem as for text length.

Recent Computational Approaches to the Measurement of SC and LD

The study of big textual material in terms of linguistic components has become possible because of the current availability of computing tools for discourse processing. Coh-Metrix, an automated tool for precise and thorough textual analysis, conveniently provides specific syntactic difficulty and lexical diversity indices (Graesser et al., 2004). Table 1 following provides a general summary of Coh-Metrix.

Table 1. Questions and answers about Coh-Metrix

Questions	Answers
What is Coh-Metrix?	Computational linguistics and recent advances in text processing technologies have lately created a large sum of complicated discourse indicators. A team at the Institute for Intelligent Systems at The University of Memphis has developed a text processing tool named Coh-Metrix that incorporates these novel and sophisticated text indices (McNamara et al., p. 164)
What function does it serve?	Coh-Metrix provides a wide number of linguistic and discourse features of a text through plentiful indices of readability, language, and cohesion. Coh-Metrix provides its textual analysis whereby automated syntactic trees and parsing, and latent semantic analysis as well as "conventional textual metrics like average sentence and word lengths and the readability formulae of Flesch Reading Ease and Flesch-Kincaid Grade Level (Klare, 1974–1975)" (McNamara, et. al., 2014).
Why should we rely on Coh-Metrix?	Syntactic complexity and lexical diversity research have started to widely benefit from Coh-Metrix for the analysis of multilevel textual features (Graesser et al., 2011) to offer subtler predictors. There has been a broad approval and employment of the tool in the related research community. The syntactic and lexical indices provided by this automated tool have been validated by several recent studies that investigated linguistic textual features as well as textual cohesion, coherence, lexical diversity, and lexical proficiency (Crossley & McNamara, 2011; Crossley et al., 2011; McNamara et al., 2010)
What are we specifically using it for? In our study, we are peculiarly interested in three syntactic complexity and two lexical diversity indices.	

Research Questions

The current study seeks to provide answers to the following research questions based on its stated objectives:

1. Is there a difference between syntactic complexity, lexical diversity, text length, and writing quality scores of learners at different curricular levels?
2. What is the relationship between syntactic complexity, lexical diversity, text length, and L2 writing quality scores assigned by human raters?
3. To what extent are syntactic complexity and lexical diversity pertinent in the perception of writing instructors who evaluate undergraduates' academic writings?

Methods

Participants

Three cohorts make up the study's participants. Table 2 shows that most participants are undergraduate and senior ELT students, with teachers and raters making up a small part of the group. The students are exposed to a variety of academic genres in spoken and written form, and they are required to produce language in the form of numerous assignments, reports, and presentations during their four-year degree program in ELT at a Turkish public university. As Wolfe-Quintero et al. (1998) wrote, "program level may be the most valid developmentally" (p. 9). We acknowledge that our first- and fourth-year students may have different levels of linguistic proficiency in light of the claim that syntactic complexity and lexical diversity in L2 writing develop over time with more instruction and exposure and vary across proficiency levels (Harley & King, 1995; Linnarud, 1986; Mazgutova & Kormos, 2015; Treffers-Daller et al., 2016; Vyatkina, 2015). Secondly, we obtained our qualitative data through semi-structured interviews with eight instructors who had been working in the same public university's four-year ELT program. The mean year of experience of instructors in teaching and assessing student writing was 16.6. Thirdly, two different scorers—one with over thirty years of expertise in teaching and grading various types of academic writing, one with over ten years of experience in teaching and assessing academic writing, and an English-native speaker who is pursuing her MA in the ELT program—evaluated the essays. A third scorer was recruited to resort to only when there was an inconsistency of 1 point or more between the two scorers.

Table 2. Distribution of participants

Participants	Number
4th Year Students	102
1st Year Students	102
Instructors	8
Raters	3
Total	215

Procedures

In this study, a sequential-explanatory mixed-method research design was employed to investigate the relationship between SC, LD, and writing quality scores in FL essays. The design was chosen to integrate quantitative and qualitative approaches in a structured manner. In the first phase, quantitative data were collected through statistical analyses, focusing on how SC, LD, and text length predicted overall writing scores assigned by human raters. Quantitative analyses included independent t-tests, correlation analyses, and hierarchical regression to examine the relationships among the variables.

In the second phase, qualitative data were gathered through semi-structured interviews with eight instructors, each with an average of 16 years of experience in rating student papers. This phase aimed to further explain the quantitative results by exploring how the instructors perceived SC and LD during the writing assessment process. The qualitative insights provided a deeper understanding of the factors influencing the scoring procedures and how instructors integrated syntactic complexity and lexical diversity into their judgments. This two-phase approach allowed for a comprehensive analysis, where the qualitative data helped to interpret and explain the patterns observed in the quantitative results.

Materials and Data Collection

We gathered an undergraduate student learner corpus as the core data of the present study. The corpus was compiled in a way that minimized the confounding effects of task (Ellis & Yuan, 2004, p. 78; Ong & Zhang, 2010; 2013) and text variables (Beers & Nagy, 2009; Halliday & Hassan, 1985; Ravid, 2005) such as genre (i.e., opinion essay) and task conditions (i.e., timed and unplanned writing within the article). Our decision-making procedure for choosing the writing topic for the opinion essay included consulting the opinions of experts via a specifically created questionnaire. This procedure aimed to immobilize the so-called topic effect. The questionnaire was comprised of 10 topics, all of which were compiled from an IELTS study recommendation page found on http://ieltsliz.com/100-ielts-essay-questions/education/ web address. The selected topics were about education, university and campus life, learning, and teaching in general. The candidate topics were presented to 20 experts who had been teaching or scoring student writing in the same department where the study was conducted. The experts were required to select the top three subjects that they believed our participants could write about with maximum ease and amount. The topic prioritized the most by 15 experts and thus selected for the current study was:

> *“University students frequently have a selection of housing options. The options available to them include living in town apartments, private student houses, or dorms on campuses. Which place would you rather live? Why? Give the rationale for your choice.”*

Afterward, the topic was placed on a writing sheet that was designed for the data collection procedure with the duration of the task, which was one hour- slightly more than a regular class hour.

The present study is based on a mixed research paradigm. Therefore, it utilizes a qualitative inquiry, as well. We profited from semi-structured interview questions to investigate the degree to which syntactic complexity and lexical diversity are involved in the perceptions of human scorers. We conducted interviews with eight instructors who had been grading students' academic papers and recorded their answers. The researcher derived the semi-structured interview questions from the related literature. After seeking expert comments throughout two feedback sessions, the questions were finally revised and given their final forms.

Data Analysis and Tools

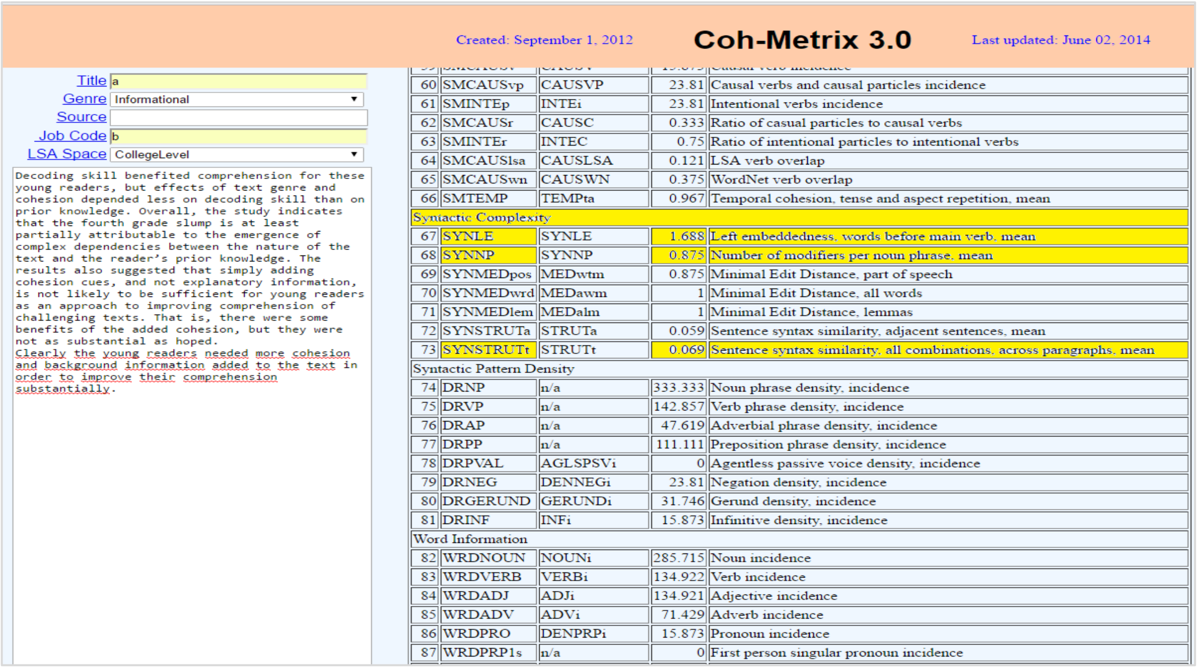
All essays were typed on Microsoft Word 2016 once they had been collected and were then processed using Coh-Metrix. The intended indices about lexical diversity and syntactic complexity were provided by Coh-Metrix. Data analysis and tools can be seen in Table 3.

The syntactic complexity and lexical diversity indices provided by Coh-Metrix as well as writing quality scores were transferred into a statistical analysis software SPSS for further analysis. The total of five intended SC and LD indices provided by the Coh-Metrix interface can be seen in the following Figures 1 and 2:

For the qualitative inquiry, procedures suggested by both Weber (1990) and Creswell (2012) were employed. Firstly, the researcher broadly read the transcribed data on several occasions by taking margin notes by hand. These margin notes, afterward, evolved into broad themes which were few. The first themes, after having been discussed for feedback with an expert, were transferred into NVivo 11 pro, which is a qualitative analysis tool for further and detailed analysis (*see* Figure 3).

Table 3. Overview of research methodology

Research Questions	Number of Participants (n)	Variables at Play	Statistical Analysis
1) Is there a statistical difference between syntactic complexity, lexical diversity, text length, and writing quality scores of learners at different curricular levels?	204	*Mean number of words before the main verb *Mean number of modifiers per noun clause *Syntactic similarity *Measure of Textual Lexical Diversity (MTLD: McCarthy and Jarvis, 2010) **VocD* (Malvern et.al., 2004). *Text length *Overall Writing Quality Scores	*Descriptive Statistics *Independent Samples T-Tests
2) What is the relationship between syntactic complexity, lexical diversity, text length, and L2 writing quality scores assigned by human raters?	204	Same as above	*Correlation Analysis *Hierarchical Regression Analysis
3) To what extent are syntactic complexity and lexical diversity pertinent in the perception of writing instructors who evaluate undergraduates’ academic writings?	8	Transcribed interviews	Content Analysis on NVivo



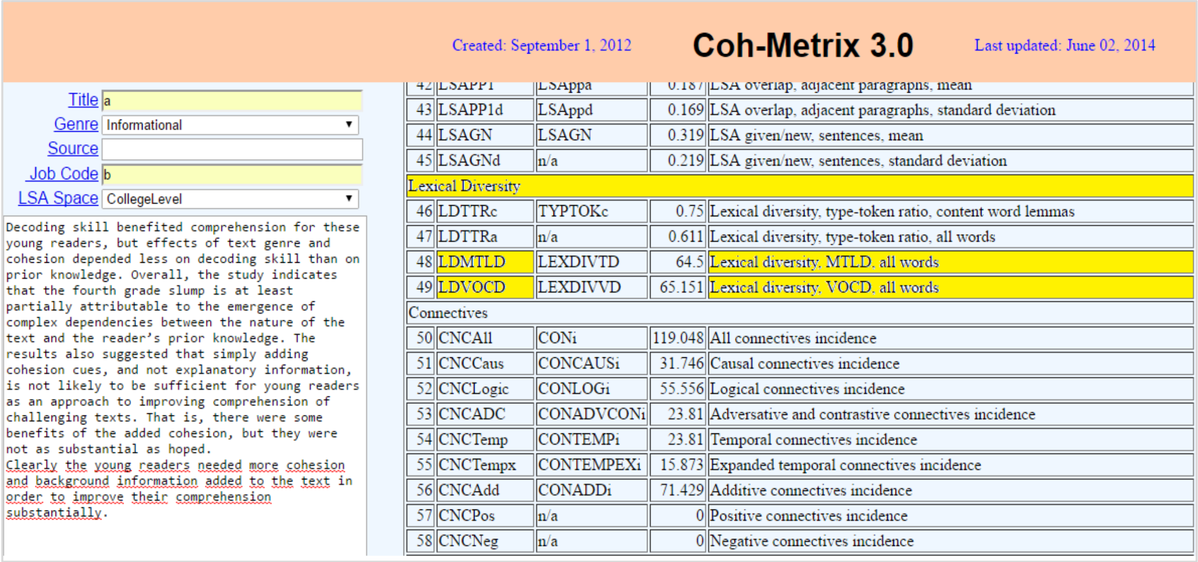
Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: June 02, 2014

Title: a
Genre: Informational
Source:
Job Code: b
LSA Space: CollegeLevel

Decoding skill benefited comprehension for these young readers, but effects of text genre and cohesion depended less on decoding skill than on prior knowledge. Overall, the study indicates that the fourth grade slump is at least partially attributable to the emergence of complex dependencies between the nature of the text and the reader’s prior knowledge. The results also suggested that simply adding cohesion cues, and not explanatory information, is not likely to be sufficient for young readers as an approach to improving comprehension of challenging texts. That is, there were some benefits of the added cohesion, but they were not as substantial as hoped.
Clearly the young readers needed more cohesion and background information added to the text in order to improve their comprehension substantially.

60	SMCAUSvp	CAUSVP	23.81	Causal verbs and causal particles incidence
61	SMINTEp	INTEi	23.81	Intentional verbs incidence
62	SMCAUSr	CAUSC	0.333	Ratio of casual particles to causal verbs
63	SMINTEr	INTEC	0.75	Ratio of intentional particles to intentional verbs
64	SMCAUSlsa	CAUSLSA	0.121	LSA verb overlap
65	SMCAUSwn	CAUSWN	0.375	WordNet verb overlap
66	SMTEMP	TEMPta	0.967	Temporal cohesion, tense and aspect repetition, mean
Syntactic Complexity				
67	SYNLE	SYNLE	1.688	Left embeddedness, words before main verb, mean
68	SYNNP	SYNNP	0.875	Number of modifiers per noun phrase, mean
69	SYNMEDpos	MEDwtm	0.875	Minimal Edit Distance, part of speech
70	SYNMEDwrd	MEDawm	1	Minimal Edit Distance, all words
71	SYNMEDlem	MEDalm	1	Minimal Edit Distance, lemmas
72	SYNSTRUTa	STRUTa	0.059	Sentence syntax similarity, adjacent sentences, mean
73	SYNSTRUTt	STRUTt	0.069	Sentence syntax similarity, all combinations, across paragraphs, mean
Syntactic Pattern Density				
74	DRNP	n/a	333.333	Noun phrase density, incidence
75	DRVP	n/a	142.857	Verb phrase density, incidence
76	DRAP	n/a	47.619	Adverbial phrase density, incidence
77	DRPP	n/a	111.111	Preposition phrase density, incidence
78	DRPVAL	AGLSPSVi	0	Agentless passive voice density, incidence
79	DRNEG	DENNEGi	23.81	Negation density, incidence
80	DRGERUND	GERUNDi	31.746	Gerund density, incidence
81	DRINF	INFi	15.873	Infinitive density, incidence
Word Information				
82	WRDNOUN	NOUNi	285.715	Noun incidence
83	WRDVERB	VERBi	134.922	Verb incidence
84	WRDADJ	ADJi	134.921	Adjective incidence
85	WRDADV	ADVi	71.429	Adverb incidence
86	WRDPRO	DENPRPi	15.873	Pronoun incidence
87	WRDPRP1s	n/a	0	First person singular pronoun incidence

Figure 1. A Coh-Metrix screenshot displaying syntactic complexity indices



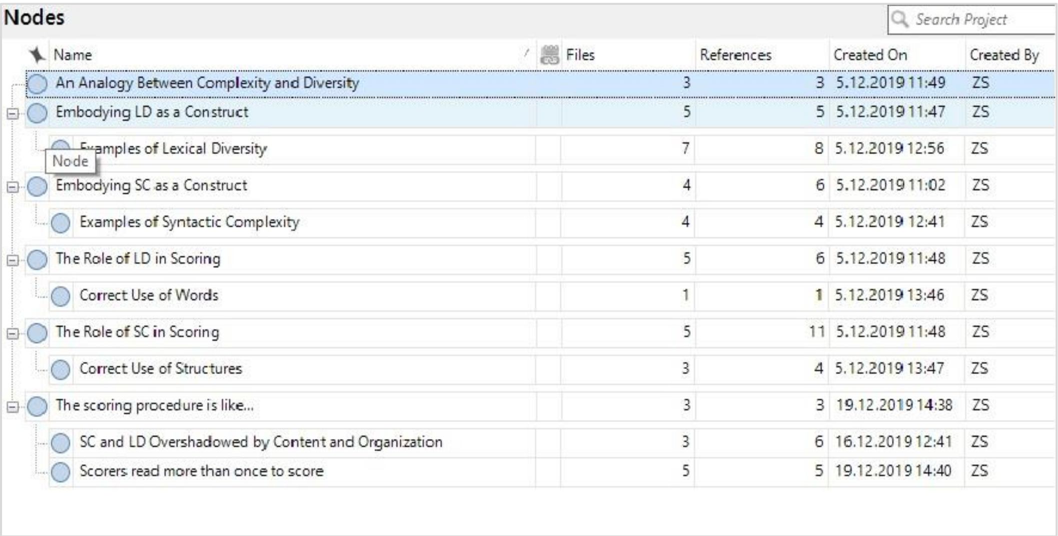
Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: June 02, 2014

Title: a
Genre: Informational
Source:
Job Code: b
LSA Space: CollegeLevel

Decoding skill benefited comprehension for these young readers, but effects of text genre and cohesion depended less on decoding skill than on prior knowledge. Overall, the study indicates that the fourth grade slump is at least partially attributable to the emergence of complex dependencies between the nature of the text and the reader’s prior knowledge. The results also suggested that simply adding cohesion cues, and not explanatory information, is not likely to be sufficient for young readers as an approach to improving comprehension of challenging texts. That is, there were some benefits of the added cohesion, but they were not as substantial as hoped.
Clearly the young readers needed more cohesion and background information added to the text in order to improve their comprehension substantially.

42	LSAPP1	LSAppa	0.187	LSA overlap, adjacent paragraphs, mean
43	LSAPP1d	LSAppd	0.169	LSA overlap, adjacent paragraphs, standard deviation
44	LSAGN	LSAGN	0.319	LSA given/new, sentences, mean
45	LSAGNd	n/a	0.219	LSA given/new, sentences, standard deviation
Lexical Diversity				
46	LDTTRc	TYPTOKc	0.75	Lexical diversity, type-token ratio, content word lemmas
47	LDTTRa	n/a	0.611	Lexical diversity, type-token ratio, all words
48	LDMTLD	LEXDIVTD	64.5	Lexical diversity, MTLD, all words
49	LDVOCD	LEXDIVVD	65.151	Lexical diversity, VOCD, all words
Connectives				
50	CNCAll	CONi	119.048	All connectives incidence
51	CNCCaus	CONCAUSi	31.746	Causal connectives incidence
52	CNCLogic	CONLOGi	55.556	Logical connectives incidence
53	CNCADC	CONADVCONi	23.81	Adversative and contrastive connectives incidence
54	CNCTemp	CONTEMPi	23.81	Temporal connectives incidence
55	CNCTempx	CONTEMPEXi	15.873	Expanded temporal connectives incidence
56	CNCAdd	CONADDi	71.429	Additive connectives incidence
57	CNCPos	n/a	0	Positive connectives incidence
58	CNCNeg	n/a	0	Negative connectives incidence

Figure 2. A Coh-Metrix screenshot displaying lexical diversity indices



Name	Files	References	Created On	Created By
An Analogy Between Complexity and Diversity	3	3	5.12.2019 11:49	ZS
Embodying LD as a Construct	5	5	5.12.2019 11:47	ZS
Examples of Lexical Diversity	7	8	5.12.2019 12:56	ZS
Embodying SC as a Construct	4	6	5.12.2019 11:02	ZS
Examples of Syntactic Complexity	4	4	5.12.2019 12:41	ZS
The Role of LD in Scoring	5	6	5.12.2019 11:48	ZS
Correct Use of Words	1	1	5.12.2019 13:46	ZS
The Role of SC in Scoring	5	11	5.12.2019 11:48	ZS
Correct Use of Structures	3	4	5.12.2019 13:47	ZS
The scoring procedure is like...	3	3	19.12.2019 14:38	ZS
SC and LD Overshadowed by Content and Organization	3	6	16.12.2019 12:41	ZS
Scorers read more than once to score	5	5	19.12.2019 14:40	ZS

Figure 3. Screenshot for the categorical themes created on NVivo

The first drawn themes were labeled as codes in NVivo and thoroughly read more than once to define persistent codes.

Overall Scoring and Intra-rater/inter-rater Reliability Check

The essays were rated by two separate raters: one with over thirty years of experience in teaching and grading various kinds of academic writing, one with over ten years of experience in teaching and assessing academic writing, and a native speaker of English who is following her MA degree in the ELT program. When there was a discrepancy of more than one point between two raters, a third rater was brought in. A standardized rubric used to score TOEFL iBT essays was employed to assess the quality of the essays (see Appendix). This rubric uses a scale of 0 to 5 to evaluate essays' overall quality, with 5 being the best possible grade.

Our scorers scored the student papers twice to ensure intra-rater reliability along with inter-rater reliability. The second scoring was carried out 6 months after the first one. The reliability check procedure was run separately across two groups of participating students. Running the Pearson product-moment correlation across scorers and scoring sessions is one method to assess the inter- and intra-rater reliability (Evans, 1996). The scorers' pseudonyms were the initial letters of their actual names (Rater Z and Rater B).

Table 4 displays both the intra-rater reliability and the inter-rater reliability across scorings for first-year students.

In terms of intra-rater reliability, though on a medium scale, only Rater Z displayed a statistically significant correlation between her scores. When it comes to inter-rater reliability scores, the two scorers – though statistically significant again- could show a weak consistency between themselves in both of the scoring procedures. Table 5 displays both the intra-rater reliability and the inter-rater reliability across scorings for the fourth-year students.

We observe some higher correlation values when it comes to the intra and inter-rater reliability values in the essays of fourth-year students. Rater Z and Rater B achieved higher correlations both within themselves and between each other in both of the scoring procedures.

Results

Curricular Level Differences Among the Investigated Variables

In this subsection, we aimed to answer the first research question of our study which was questioning whether there was a difference between text length, overall writing quality, syntactic complexity, and lexical diversity scores of learners at different curricular levels (e.g. 1st and 4th year students).

Differences in Text Length

An independent samples t-test was run to find out if the mean differences in word count between groups were statistically significant or not. In Table 6, the t-test finding showed that 4th-year students' essays (M=361,38; SD=113,7) contain more words than 1st-year students' essays (M=280,86; SD=71,6) and that this mean difference is statistically significant *t* (202) =6,048, *p*=.000.

Differences in Overall Writing Scores

An independent samples t-test was conducted to see if there is a significant difference between 1st year and 4th-year students' writing quality scores. There was a significant difference between the means of 1st-year students' writing quality scores (M=3.2, SD=.3.20) and 4th-year students' writing quality scores (M=3,7, SD=.619) as displayed in Table 7.

Table 4. Results of Pearson correlations coefficients between two raters across two scoring procedures (for 1st-year students' scores)

	Rater Z 1st Scoring	Rater Z 2nd Scoring	Rater B 1st Scoring	Rater B 2nd scoring
Rater Z 1st Scoring	1			
Rater Z 2nd Scoring	.449**	1		
Rater B 1st Scoring	.342**	.198*	1	
Rater B 2nd Scoring	.003	.047	.132	1

** Correlation is significant at the 0.01 level (2-tailed).

Table 5. Results of Pearson correlations coefficients between two raters across two scoring procedures (for 4th-year students' scores)

	Rater Z 1st Scoring	Rater Z 2nd Scoring	Rater B 1st Scoring	Rater B 2nd scoring
Rater Z 1st Scoring	1			
Rater Z 2nd Scoring	.509**	1		
Rater B 1st Scoring	.546**	.346**	1	
Rater B 2nd Scoring	.331**	.316**	.469**	1

Table 6. A numerical comparison of 1st and 4th year students' essays

Curricular Level	n	Total Word Count	Min.	Max.	M	Std. Deviation
1st Year Students	102	28.648	113	473	281	71.619
4th Year Students	102	36.861	127	685	361	113.792

Table 7. Results of independent samples t-test for writing quality scores by curricular level

	1st Year Students			4th Year Students					
	M	SD	n	M	SD	n	t	df	p
Writing Quality in the First Scoring	3.2	3.20	102	3.7	.619	102	-9.95	202	.000
Writing Quality in the Second Scoring	3.5	.329	102	3.8	.344	102	-6.66	202	.000

Table 8. Results of independent samples t-test for '3 SC indices of Coh-Metrix' by curricular level

	1st Year Students			4th Year Students					
	M	**SD**	**n**	**M**	**SD**	**n**	**t**	**df**	**p**
Number of words before the main verb	3.76	.123	102	4.15	.124	102	-2.24	202	<.05
Mean Number of Modifiers per Noun Phrase	.577	.121	102	.636	.136	102	-3.25	202	<.001
Syntactic Similarity	.111	.028	102	.121	.033	102	-2.36	202	<.05

Table 9. Results of independent samples t-test for '2 LD indices' by curricular level

	1st Year Students			4th Year Students					
	M	**SD**	**n**	**M**	**SD**	**n**	**t**	**df**	**p**
MTLD	68.37	14.27	102	71.51	15.67	102	-1.49	202	>.05
VocD	75.70	15.46	102	80.59	14.89	102	-2.23	202	<.05

Specifically, these results suggest that our 4th-year students scored higher than the 1st-year students and this difference in the mean scores was found to be statistically significant [t (202) =-9.957, p=.000)].

Differences in Syntactic Complexity

In independent samples t-test results, we found a mean difference in the number of 'words coming before the main verb' in each sentence of the compositions of 1st (M=3.76 SD=1,24) and 4th-year students (M=4,15 SD=1,26), these mean differences are statistically significant as seen in Table 8.

Second, Coh-Metrix provides noun phrase (NP) density and the mean 'number of modifiers per NP' as a syntactic complexity index. A statistically significant mean difference was found in this index as well. The fourth-year students used a higher number of modifiers per NP than 1st-year students and this difference was found statistically significant (p<.001). Third, the mean scores of 1st (M=.111 SD=.028) and 4th-year students (M=.121 SD=.033) in the 'syntactic similarity' index of Coh-Metrix were also different; these mean differences were found to be statistically significant. However, Coh-Metrix measures syntactic similarity differently from the other two syntactic complexity indices, which is important to notice. In other words, the lower the number, and less comparable the structures are, indicating a broader variety of syntactic structures used in an essay.

Differences in Lexical Diversity

We used lexical diversity indices reported by Coh-Metrix, which are more sophisticated, and reliable than traditional measures like TTR and free from text length effect. They, namely, are the Measure of Textual Lexical Diversity (MTLD: McCarthy and Jarvis, 2010) and VocD (Malvern et al., 2004).

We found statistically significant differences within two measures for lexical diversity between compositions written by students in the first year and those in the fourth year. The 4th-year students outperformed the 1st-year students in both indices based on different mean scores, but only in VocD was the difference statistically significant as shown in Table 9.

Correlations of Syntactic Complexity, Lexical Diversity, and Text Length with Writing Quality and Variances Explained

In this section, we aim to answer our 2nd research question which was about the relationships of syntactic complexity, lexical diversity, and text length with writing quality scores. We computed a Pearson product-moment correlation coefficient to examine the relationship between Syntactic Similarity and the Number of Modifiers, Number of words before a Main verb, MTLD, VocD, Text Length, and Writing Quality. The value is displayed in Table 10. Text length, though moderate but on a statistically significant scale, showed the highest positive correlation with writing quality. As comes to SC and LD, only two modifiers per NP2 and VocD could yield weak but statistically significant positive correlations with the dependent variable. It is noticeable that human scorers could not grasp subtle details related to the complexity and diversity of a text from a syntax or vocabulary perspective, which is discussed more in detail in the Discussion part.

Another noteworthy finding of the correlation test is that LD indices positively correlated with each other implying that these are valid and reliable indices. The same goes for SC indices as well with one subtle difference. The 'Syntactic similarity' index negatively correlated with other Coh-Metrix indices since it worked peculiarly. Higher scores of 'syntactic similarities' are a sign of repeating patterns of syntax and repetitive vocabulary in contrast to complexity and diversity notions.

Table 10. Results of Pearson correlations coefficients among seven variables

	Syntactic Similarity	**Number of Modifiers**	**Number of words before the Main verb**	**MTLD**	**VocD**	**Text Length**	**Writing Quality**
Syntactic Similarity	1						
Number of Modifiers	-.219**	1					
Number of words before the Main verb	-.417**	.383**	1				
MTLD	-.222**	.246**	.223**	1			
VocD	-.139*	.192**	.156*	.815**	1		
Text Length	.038	.155*	.170*	.011	.053	1	
Writing Quality	.092	.141*	.110	.088	.177*	.449**	1

*** Correlation is significant at the 0.01 level (2-tailed)*

**Correlation is significant at the 0.05 level (2-tailed)*

Table 11. Hierarchical multiple regression analysis with a three-layered model (dependent variable; writing quality overall scores)

Model	R	R Square	Standard Error	F Model	R Square Change	F Change
Text Length	.449	.202	.422	51.02*	.202	51.02*
Lexical Diversity Indices	.480	.230	.416	19.95**	.029	3.72**
Syntactic Complexity Indices	.495	.245	.415	10.66**	.015	1.28**

Variances Explained in Writing Quality Scores

Hierarchical Multiple Regression Analysis with a Three-Layered Model is displayed in Table 11. The hierarchical multiple regression analysis regarding the factors that influence students' overall writing scores. It indicates that text length by itself was a significant predictor, accounting for 20.2% of the variance in the writing scores, meaning it had a strong impact on how the scores varied. In contrast, when looking at SC and LD together, they could only explain 4.4% of the variance, suggesting that while these linguistic features do contribute to the scores, their combined influence was much smaller compared to text length. Essentially, text length was the most significant factor in determining writing quality, while SC and LD played a more limited role.

Unfolding SC and LD: Embodying them as a Construct

In this section, we responded to the third research question, which was designed to explore how ELT teachers view LD and SC as they evaluate undergraduates' academic writing. The first theme focuses on how eight instructors who have been grading essays for students for an average of 16 years have conceptualized SC and LD.

> *Complex writing is associated with the use of various clauses (adverbial, adjective, and noun clauses) and conjunctions. Simple sentences following a subject-verb-object order are considered elementary, and a lack of these complex structures in writing diminishes its sophistication. (Inst.5)*
>
> *Lexical diversity, including synonyms, antonyms, idiomatic expressions, and chunks, is vital in writing. In tasks like cause-and-effect essays, students should avoid repetitive phrases like "first cause" and use varied terms such as "impact" or "influence" to display linguistic sophistication. (Inst.1)*

Providing structural variety is one of the most prominent features that have been associated with SC. Likewise, the repetitive and frequent word use is seen as contrary to LD since, as the name implies, lexical diversity is closely related to the wide range of words, both in meaning and number. When it comes to the examples or signs of SC and LD in a text, the instructors regard using passive structures and prepositions correctly as well as embedded structures and inversions are among the patterns that signal SC;

> *Using advanced-level vocabulary, especially noun forms of verbs, is viewed as a mark of higher proficiency. Less common vocabulary elevates the perceived quality of the writing. (Inst.8)*

In the interviews, noun forms, synonyms, antonyms, phrasal verbs, and collocations were counted as the patterns of lexical usage that point out LD;

> *I can say [lexical diversity is in the writings] which consists of advanced level vocabulary and perhaps noun forms. Noun forms of most verbs are accepted as more advanced. Therefore, [the use of less common vocabulary] (Inst.2)*

The interviewees' comments highlight that varied use of both syntax and vocabulary should be accurate and appropriate for the task. The task requirements and linguistic accuracy, for the sake of SC and LD, should not be given up.

The Role of SC and LD in Scoring the Students' Essays

In our analysis, we focused on the question of how syntactic complexity perceived by our instructors affects their scores. We already reported what kind of constructions would evoke syntactic complexity in our instructors' minds. Two instructors discuss their high expectations for English language teaching majors, stating that these students should have a high degree of language ability and show this proficiency in their writing by using syntactically complicated sentences. According to these professors, employing solely straightforward yet true statements will not result in good marks:

> *We are telling them [our students], 'You are going to be English teachers.' So, there must be a level of mastery. They must show us that they can use different forms and structures. If you are only using simple sentences, even if they are grammatically correct, you may not get high grades. I expect that complexity. (Inst.3)*
>
> *If sentences are accurate but simple, they cannot get high scores because what I expect from an ELT student is not simplicity (Inst.6)*

Similarly, according to an instructor, in addition to affecting the language score of the writing, a large variety of vocabulary can reflect a vast variety of writing idea units:

> *[Lexical diversity] I think affects the language score. If the language is correct, of course, this will positively affect the content. There's a difference between a student always saying 'thing or cause' to express an idea. However, if they use 'thing' sometimes and 'cause' at others, this variety, I suppose, affects the content score. (Inst.2)*

Inst 2 above stated the close relation of linguistic and rhetorical features of a text to be evaluated while saying that LD could also affect the content of the writing. In that verbatim quotation and elsewhere, we witnessed that content and organizational patterns may overshadow SD and LD. Our instructors, as they reported in interviews, make an order of importance on their minds while reading student papers and in this order of importance, content and organization come first, leaving SC and LD behind. One instructor (Inst. 7) said that "*an essay written with a good command of English can make me suppose that the content is also well developed, thus at the very beginning I divide these dimensions from each other*". The below given verbatim quotations exemplify the point;

> *I generally start scoring the content. I love scoring with a focus on content and organization. Because the mechanic part of the writing can affect me negatively (Inst. 7)*
>
> *First of all, I look into the content. And then I look into our expectations. For example, is there what there should be in an opinion essay? I review again like this and lastly, I look into grammar, spelling, and punctuation. (Inst. 5)*

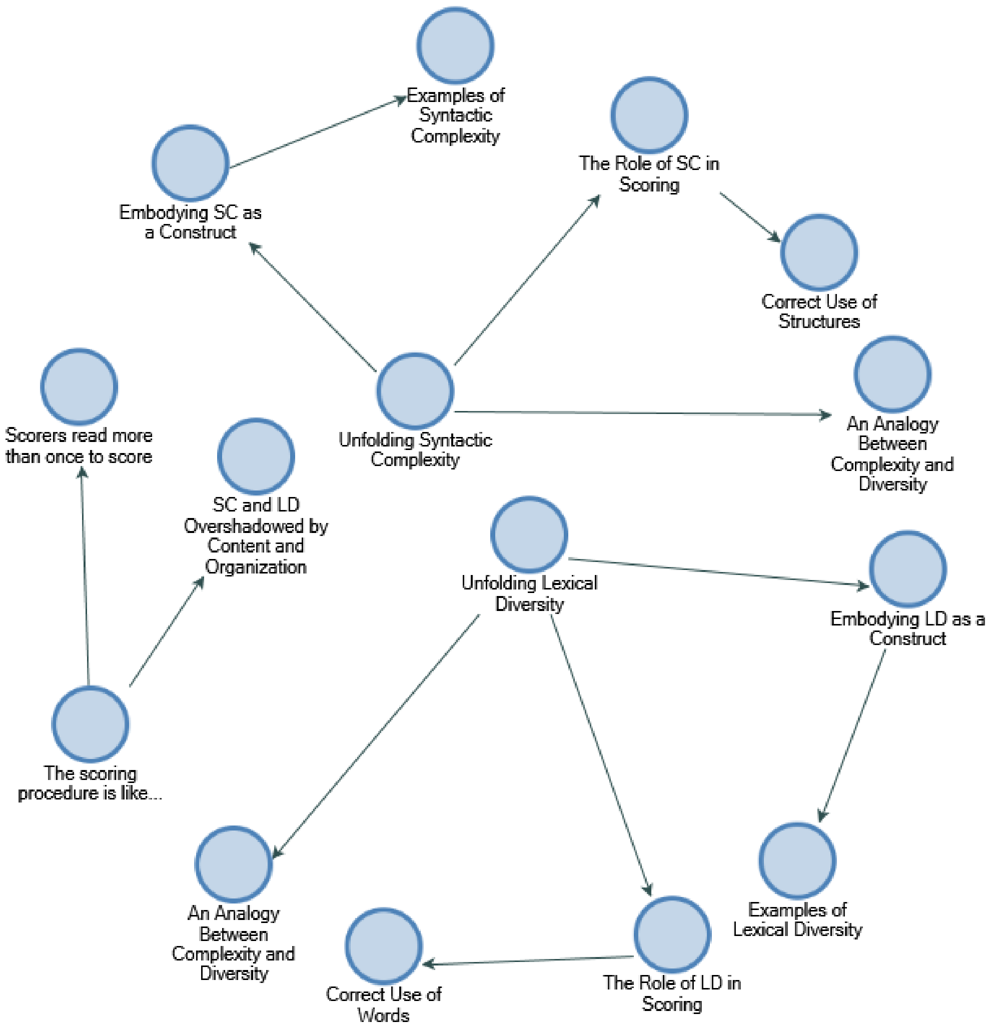

Examples of Syntactic Complexity
The Role of SC in Scoring
Embodying SC as a Construct
Correct Use of Structures
An Analogy Between Complexity and Diversity
Scorers read more than once to score
Unfolding Syntactic Complexity
SC and LD Overshadowed by Content and Organization
Unfolding Lexical Diversity
Embodying LD as a Construct
The scoring procedure is like...
An Analogy Between Complexity and Diversity
Examples of Lexical Diversity
Correct Use of Words
The Role of LD in Scoring

Figure 4. The Thematic Display of 'Qualitative Results Summary'

Figure 4 displays an overview of the qualitative findings of the present study. The findings include browsing the contents of SC and LD as well as their examples in a given text and their contribution to the overall writing score given by human judges. The results also highlight the perceived significance of content and organization in the process of assessment of student writings in foreign languages. Both quantitative and qualitative findings of the study will be discussed in depth in the following chapter.

Discussion

Issues of Syntactic Complexity with Regards to Scoring

One of the current study's most notable conclusions is that as students' compositions get syntactically more complex, their general language competency increases. This finding is consistent with several other studies that found better writers might produce more complicated works of writing with greater exposure to and practice with the language (Johansson & Geisler, 2011; Mazgutova & Kormos, 2015; Norris & Ortega, 2009; Stockwell & Harrington, 2003; Stockwell, 2005; Vyatkina et al., 2015). It has been proposed in the literature that learning more advanced and specific grammatical constructions might help learners come up with novel terms and complex ideas (Beers & Nagy, 2009). As for the relationship of syntactic complexity and writing quality scores, our study which was carried out in a FL context could only pose weak correlations between syntactic complexity and writing quality. This finding contradicts several previous studies in the literature. On the other hand, we should remember that comparing the studies on complexity issues needs much attention partly due to a lack of uniformity in the complexity measures and more importantly due to the lack of a clear definition of the complexity construct (Bulte & Housen, 2014).

In line with our study, "nominalizations, attributive adjectives, and prepositional phrases" (Beers & Nagy, 2009, p. 187) were found to be visible in evaluating the syntactic complexity of written pieces. Likewise, we also found -though very weak- a positive correlation between the number of modifiers (as an index of SC) and writing quality.

In a seminal work of research synthesis, Ortega (2003) concluded that in syntactic complexity and writing relationship research which was conducted in ESL settings, participants generated more complex writings compared to those in the studies conducted in EFL instructional settings One reason for this could be the differences between EFL and ESL instructional settings. As suggested by Ortega (2003), in EFL learning environments learners might not have the experience of learning a language as in ESL settings, which may be hindering the fast development of learners in FL settings. Another reason for the weak correlations between syntactic complexity and writing quality might be the individual beliefs and approaches of human scorers to complexity in writing. As can be understood from our participating scorers' remarks, some demand and seek syntactic complexity from their students while some do not and value simplicity and accuracy more. Moreover, general impressions of human scorers, even if they follow a standardized criterion, are more prone to detect some organizational and content issues in writing. Human scorers might be overlooking the details and delicate signs of syntactic complexity. On the other hand, automated text processing tools like Coh-Metrix in our case can well detect and calculate syntactic complexity in a computerized certainty. Therefore, it is important to emphasize that the weak and low correlations are between the overall scores given by human raters and the individual indices generated by a computerized text analysis tool. In addition, human scorers might have different expectations from their students' writings in terms of the number and nature of examples given or the genre-specific rules to be followed. Whereas automated text processing tools do not hold any judgments or expectations, but rather only calculate syntactic complexity based on several pre-ordered indices.

Issues of Lexical Diversity with Regards to Scoring

Regarding methods for comprehending and defining lexical diversity, our study proposed, in light of qualitative data, that it consists essentially of employing as many unique and obscure terms as feasible in FL student writing. The key to comprehending lexical diversity was discovered to be the number of words that are present in a student text. A large body of prior research supported this finding. To date, lexical diversity has been referred to by several names, including "lexical variation" (Engber, 1995), "lexical density" (O'Loughlin, 1995), "a combination of lexical variation and lexical sophistication" (Laufer, 2003, p. 24), and "lexical richness" as coined by (Daller et al., 2003). Overall, the number of words is what determines these various characterizations.

Some earlier studies that compared the lexical diversity of written texts were conducted between native and non-native groups of English learners (Harley & King, 1989; Linnarud, 1986). Others were conducted in short-term (Bulte & Housen, 2014) or long-term (Mazgutova & Kormos, 2015) ESL language programs and with learners of English of different L1 backgrounds (Jarvis, 2002; Yu, 2009). In all of these studies, lexical diversity was found to be developing over time and with more exposure to language through instruction. Likewise, our study produced similar findings in that our 4th-year students wrote essays that were lexically more diverse than those of our 1st-year students.

As for the lexical diversity and its relationship with overall FL writing quality scores, our study showed only a weak and positive correlation, though statistically significant, with the Vocab-D measure of lexical diversity and overall quality scores. The other index of lexical diversity (MTLD) could not yield any statistically significant correlation. These findings accorded with several previous studies. In the literature, some studies produced statistically significant and positive correlations between LD and FL writing quality as well as studies that did not. For example, as for predicting overall writing quality, the D-value exerted a weak and non-significant correlation in Bulte and Housen (2014). Likewise, Engber (1995) also put forward a non-significant and low correlation with writing quality scores (r=.23), which means that the "percentage of lexical words has little, if any, relationship to quality" (p. 148). Similarly, in a study with English learners of different L1 backgrounds, Jarvis (2002) presented, though moderate, a significant and positive correlation only between Swedish students' lexical diversity and writing scores. The same study, however, showed statistically non-significant and low correlations between lexical diversity and writing scores of American and Finnish students.

There were, of course, previous studies which contradicted our findings. In other words, several studies found a positive and moderate or strong correlation between lexical diversity and writing quality scores. However, the methodology of each research study was different. For example, Crossley et.al. (2010) broadly characterized lexical diversity as a knowledge "breadth of lexical knowledge, depth of lexical knowledge and the accessibility to core lexical items" (p. 1). These three broad categories were measured through 10 different incidences provided by Coh-Metrix and the findings produced a positive correlation (r=.66) between these broad categories of lexical knowledge and writing quality scores assigned to 240 essays.

Issues of Text Length with Regards to Scoring

Text length in our study was the variable that produced the strongest correlation with human scorers. We found a moderately strong and positive correlation which was statistically significant between text length and writing quality scores assigned by human raters. This finding is likely since text length is comparatively easier for human scorers to detect and evaluate. As our participating scorers stated, scorers might read the student essays more than once to evaluate it from several respects and one of these respects could be the text length since it can be caught even with a glimpse of eye. Similarly, Jarvis et al. (2003) found that text length positively correlated with all 21 linguistic features of 160 ESL and 178 EFL student essays which were assigned high scores by human raters.

Text length has been strongly associated with evaluation and writing quality. Text length also showed up in our study as a major variable that affected participants' writing quality scores. However, contrary to our findings, some research found that more proficient learners could pack more complex ideas into smaller sentences, thus producing smaller or shorter texts (Becker, 2010). On the other hand, Bi and Jiang (2020) rather more recently considered text length as an indicator of syntactic complexity and found out that text length together with complex nominals per clause, and clauses per T-unit as the best predictors of human judgments of 410 narratives of Chinese EFL learners. Therefore, it is possible to claim that text length in terms of syntactic complexity has an ambiguous nature as in our study we found a moderate positive correlation between text length and writing quality scores.

Text length, in our study as a confounding variable, also explained the variance in writing quality scores on a significant scale. Both alone and together with SC and LD on the three-faceted model, text length explained 20% and 24% of the variance respectively. Mellor (2011) also yielded similar findings in his study. Mellor (2011) wrote that "lexical diversity together with text length can more accurately predict essay quality than either feature alone in this set of essays" (Mellor, 2011, p. 9). Essay length, however, was found superior over lexical diversity indices in predicting essay quality.

Conclusion and Implications for EFL Writing Pedagogy and Future Research

This study investigated the relationship between SC, LD, TL, and writing quality in the context of FL writing among pre-service teachers. The results indicated that text length was the strongest predictor of writing quality scores, followed by modest contributions from SC and LD. Fourth-year students significantly outperformed first-year students in all the examined indices, suggesting that linguistic features develop with greater exposure and instruction. However, qualitative data from the interviews revealed that human raters varied in their awareness and prioritization of SC and LD, with some instructors emphasizing content and organization over linguistic features.

The findings suggest several implications for EFL writing pedagogy. First, the emphasis on SC and LD in writing instruction should be balanced with training that enhances students' overall organization and content-generation skills. The weak correlations between SC/LD and writing quality scores imply that while linguistic complexity contributes to writing quality, it is not the sole determinant. Therefore, writing pedagogy should not only focus on enhancing syntactic and lexical features but also on ensuring that students can organize and articulate their ideas effectively.

Moreover, the study highlights the need for standardized writing assessment practices in EFL contexts, where human raters may place different emphases on linguistic versus content-related features. Incorporating automated tools like Coh-Metrix into the assessment process could help reduce subjective variations and provide a more consistent evaluation framework.

Future research should expand the scope of this study by exploring the relationship between SC, LD, and writing quality across different genres and proficiency levels. Longitudinal studies tracking the development of these linguistic features over time would provide further insights into how SC and LD evolve with instruction. Additionally, integrating more comprehensive qualitative measures to examine how instructors' perceptions of writing complexity influence their scoring would enrich the understanding of human judgment in FL writing assessments.

Several limitations must be acknowledged. First, the study's focus on pre-service teachers in a single institutional context limits the generalizability of the findings. Replicating the study across different educational settings and with a more diverse participant pool would provide more robust conclusions. Second, the use of Coh-Metrix, while beneficial for measuring SC and LD, does not capture the full complexity of human judgment in writing assessments. Lastly, the cross-sectional design of the study restricts our ability to track how students' writing skills develop over time, necessitating future longitudinal research.

Author Contributions

All authors took an equal part in all processes of the article. All authors have read and approved the final version of the study.

Ethical Declaration

Anadolu University Ethics Committee granted approval for the present research on 29.04.2019 (Protocol No. 31064).

Conflict of Interest

The authors declare that there is no conflict of interest with any institution or person within the scope of the study.

References

Becker, A. (2010). Distinguishing linguistic and discourse features in ESL students' written performance. *Modern Journal of Applied Linguistics*, *2*, 406-424.

Beers, S. F., and Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing*, *22*(2), 185-200.

Bi, P., and Jiang, J. (2020). Syntactic complexity in assessing young adolescent EFL learners' writings: Syntactic elaboration and diversity. *System*, *91*, 102248. doi.org/10.1016/j.system.2020.102248

Bulté, B., and Housen, A. (2012). Defining and operationalizing L2 complexity. Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA, *32*, 21.

Bulté, B., and Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65. doi.org/10.1016/j.jslw.2014.09.005

Casal, J. E., and Lu, X. (2021). 'Maybe complicated is a better word': Second-language English graduate student responses to syntactic complexity in a genre-based academic writing course. *International Journal of English for Academic Purposes: Research and Practice*, *2021*(Spring), 95-115.

Creswell, J. W. (2002). Educational research: Planning, conducting, and evaluating quantitative. Prentice Hall

Crossley, S. A., and McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson and R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., and McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyzes first and second-language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(2-3), 170-191. doi.org/10.1504/IJCEELL.2011.040197

Daller, H., R. Van Hout, and J. Treffers-Daller. (2003). 'Lexical richness in the spontaneous speech of bilinguals,' *Applied Linguistics* 24: 197–222. doi.org/10.1093/applin/24.2.197

Evans, J.D. (1996) Straightforward statistics for the behavioral sciences. *Pacific Grove, CA: Brooks/Cole Publishing*

Foster, P., and Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*(3), 299-323. doi.org/10.1017/S0272263100015047

Geiser, S., and Studley, R. (2001, October). *UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Paper presented at the Meeting of the Board of Admissions and Relations with Schools of the University of California.

Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234 doi.org/10.3102/0013189X11413

Guo, L., Crossley, S. A., and McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218-238. doi.org/10.1016/j.asw.2013.05.002

Halliday, M. A. K., and Hassan, R. (1985). Language, context, and text: Aspect of language in a social semiotic perspective. *Geelong, Australia: Deakin University Press.*

Harley, B., and King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition, 11*, 415-439. https://doi.org/10.1017/S0272263100008421

Hmelo, C., Holton, D., and Kolodner, J. (2000). Designing to Learn About Complex Systems. *Journal of the Learning Sciences*, 9, 247 - 298. https://doi.org/10.1207/S15327809JLS0903_2.

Jarvis, S. (2002). Short texts, best-fitting curves, and new measures of lexical diversity. *Language Testing*, *19*(1), 57-84. doi.org/10.1191/0265532202lt220oa

Jarvis, S., Grant, L., Bikowski, D., and Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377-403. doi.org/10.1016/j.jslw.2003.09.001

Jenkins, J. R., Johnson, E., and Hileman, J. (2004). When is reading also writing: Sources of individual differences on the new reading performance assessments. Scientific Studies of Reading, 8(2), 125-151. doi.org/10.1207/s1532799xssr0802_2

Johansson, C., and Geisler, C. (2011). Syntactic aspects of the writing of Swedish L2 learners of English. *Language and Computers-Studies in Practical Linguistics*, *73*(1), 139. doi.org/10.1163/9789401206884_009

Laufer, B. and P. Nation. (1995). 'Vocabulary size and use – lexical richness in L2 written production,' *Applied Linguistics* 16: 307–22. doi.org/10.1093/applin/16.3.307

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning*, *40*(3), 387-417. doi.org/10.1111/j.1467-1770.1990.tb00669.x

Linnarud, M. (1986). Lexis in composition: A performance analysis of Swedish learners' written *English. Malmo, Sweden: Liber Forlag Malmo.*

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

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

Malvern, D. D., Richards, B. J., Chipere, N., and Durán, P. (2004). Lexical diversity and language development. *Houndmills, Hampshire, UK: Palgrave Macmillan*.

Mazgutova, D., and Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes program. *Journal of Second Language Writing*, *29*, 3-15. doi.org/10.1016/j.jslw.2015.06.004

McNamara, D. S., Crossley, S. A., and McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*, 57-86.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. *Cambridge University Press.*

Muter, V., Hulme, C., Snowling, M., and Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Developmental psychology*, 40 5, 665-81. https://doi.org/10.1037/0012-1649.40.5.665.

Norris, J. M., and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics*, 30, 555–578. doi.org/10.1093/applin/amp044

Olinghouse, N. G. and J. Wilson. (2013). 'The relationship between vocabulary and writing quality in three genres,' *Reading and Writing* 26: 45–65

Ong, J., and Zhang, L. J. (2010). Effects of task complexity on fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, 19, 218–233. doi.org/10.1016/j.jslw.2010.10.003

Ong, J., and Zhang, L. J. (2013). Effects of manipulation of cognitive processes on English-as-a-foreign-language (EFL) writers' text quality. *TESOL Quarterly*, 47, 375–398. doi.org/10.1002/tesq.55

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24, 492–518 doi.org/10.1093/applin/24.4.492

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann and B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). Berlin: De Gruyter.

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82-94. doi.org/10.1016/j.jslw.2015.06.008

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(1), 117-134. doi.org/10.1177/0267658314536435

Ravid, D. (2005). Emergence of linguistic complexity in later language development: Evidence from expository text construction. *In Perspectives on language and language development* (pp. 337-355). Springer, Boston, MA.

Reading, *8*, 125-151 doi.org/10.1207/s1532799xssr0802_2

Stockwell, G. and Harrington, M. (2003). The incidental development of L2 proficiency in NS-NNS email interactions. *CALICO Journal* 20(2): 337–359.

Stockwell, G. (2005). Syntactical and lexical development in NNS-NNS asynchronous CMC. *The JALT CALL Journal*, *1*(3), 33-49.

Treffers-Daller, J., Parslow, P., and Williams, S. (2016). Back to basics: how measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302-327. doi.org/10.1093/applin/amw009

Vyatkina, N., Hirschmann, H., and Golcher, F. (2015). Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing*, *29*, 28-50. doi.org/10.1016/j.jslw.2015.06.006

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). *Thousand Oaks*, CA: Sage Publications

Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity. *Honolulu, HI: University of Hawaii Press.*

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236-259. doi.org/10.1093/applin/amp024