languages

MDPI

Article

Automated Discourse Analysis Techniques and Implications for Writing Assessment

Trisevgeni Liontou

Department of English Language and Literature, School of Philosophy, National and Kapodistrian University of Athens, 157 84 Athens, Greece; tliontou@enl.uoa.gr

Abstract: Analysing writing development as a function of foreign language competence is important in secondary school children because the developmental patterns are strongest at a young age when successful interventions are needed. Although a number of researchers have explored the degree to which specific textual characteristics in EFL students' essays are associated with high and low ratings by teachers, the extent to which such characteristics are associated with rater-mediated assessment under standard exam conditions remains relatively unexplored. Motivated by the above void in pertinent literature, the overall aim of the present study was to investigate the relationship between specific discourse features present in the writing scripts of EFL learners sitting for the British Council's APTIS for TEENS exam and the assigned scores during operational scoring by specially trained raters. A total of 800 international EFL students aged 13 to 15 years old took part in the study, and 800 scored written essays on the same task prompt of the pertinent test produced under standard exam conditions were analysed. The results showed statistically significant differences ($p \leq 0.05$) between the linguistic features identified in the essays produced by young EFL learners at different levels of language competence. The main text features that were repeatedly found to make a significant contribution to distinguishing scores assigned to texts both within and across levels were word frequency, word abstractness, lexical diversity, lexical and semantic overlap, all of which could be used to obtain a numerical cut-off point between proficiency levels. These findings support the notion that progress in L2 writing is primarily associated with producing more elaborate texts with more sophisticated words, more complex sentence structure and fewer cohesive features as a function of increased language competence. The findings of the study could provide practical guidance to EFL teachers, material developers and test designers as to the kind of linguistic strategies young EFL learners develop as a function of their level of language competence and suggestions to consider when designing EFL classroom curricula, writing skills textbooks and exam papers on written production.

Keywords: text complexity; essays; written production; English language competence; writing tasks; rating criteria; assessing writing

Citation: Liontou, Trisevgeni. 2023. Automated Discourse Analysis Techniques and Implications for Writing Assessment. *Languages* 8: 3. https://doi.org/10.3390/languages8010003

Academic Editors: Dina Tsagari, Henrik Bøhn, Juana M. Liceras and Raquel Fernández Fuertes

Received: 7 November 2021
Revised: 20 July 2022
Accepted: 30 September 2022
Published: 21 December 2022

1. Introduction

Analysing foreign language writing development is important in elementary and secondary school children because the developmental patterns are strongest at a young age when successful interventions are needed (Berninger et al. 1994; Haswell 2000; McNamara et al. 2010b; Muncie 2002; Witte and Faigley 1981). Over the past thirty years, such an exploration across levels of foreign language competence has provided researchers with crucial information about how writing skills change as cognitive and linguistic functions develop (Crossley et al. 2008, 2011, 2016; Ferris 1994). To begin with, Ferris (1994) gathered a corpus of 160 essays on one specific task (i.e., culture shock) written by EFL students as part of a university placement examination, and text analysis revealed that freshman university students at higher levels of L2 proficiency tended to make use of a wider range of lexical items along with more complex syntactic constructions and a variety of referential

cohesion devices (i.e., synonymy, antonymy and deictic reference). These findings are in accord with a more recent study by Espada-Gustilo (2011) who adopted a corpus linguistic approach to analyse essays produced by first-year college EFL students in the Philippines. The analyses of the 150 collected scripts revealed that most of the general text characteristics (e.g., fluency, unique words, word per sentence, number of sentences and paragraphs) and lexical (except conjunctions), clause-level and grammatical features all had a steady increase across proficiency levels, which could be treated as an indication that writers of highly scored essays were likely to employ more of these structures, while at the same time, the presence of such structures seemed to impact raters' assignment of scores on those essays.

In an older study, Crossley and McNamara (2009) used Coh-Metrix and discriminant function analysis to investigate how cohesion and lexical networks can distinguish texts written by L1 or L2 writers of English. Data analysis revealed that L2 writers produced texts with less lexical variety, specificity and sophistication than their L1 counterparts. In addition, they created texts that were less abstract and ambiguous, depended less on meaningful words, introduced too many words and ideas and were less able to provide lexical overlap between propositions than L1 writers, thus providing less causal and spatial cohesion. At the same time, L2 writers used more locational nouns and frequent words than L1 writers. Based on these findings, the authors supported the view that the results of their study could have implications on teaching, as L2 learners tend to create less coherent and comprehensible writing than their L1 counterparts. Following that study, Crossley et al. (2011) attempted to provide a model which predicts and explains human judgments of L2 lexical proficiency. Human raters evaluated 240 texts, which were then correlated to a variety of lexical indices provided from the computational tool Coh-Metrix. The results showed that three lexical indices provided by Coh-Metrix, namely lexical diversity, word hypernymy and word frequency, were sufficient to predict 44% of the variance of the human evaluations of lexical proficiency; thus human evaluations and lexical indices had strong relationships.

In another study, Crossley and McNamara (2014) examined second language (L2) syntactic development in conjunction with the effects such development has on human judgments of writing quality (i.e., judgments of both overall writing proficiency and more fine-grained judgments of syntactic proficiency). Essays collected from 57 EFL learners in an academic context were analysed for growth and scoring patterns using syntactic complexity indices calculated once again using the computational tool Coh-Metrix. The analyses demonstrated that significant growth in syntactic complexity occurred in the L2 writers as a function of time spent studying English. However, only one of the syntactic features that demonstrated growth in the L2 learners was also predictive of human judgments of L2 writing quality. Interpretation of the findings suggest that over the course of a semester L2 writers produced texts that were increasingly aligned with academic writing (i.e., texts that contained more nouns and phrasal complexity) but that human raters assessed text quality based on structures aligned with spoken discourse (i.e., clausal complexity). A year later, Yang et al. (2015) compiled a corpus of 380 essays by graduate students and examined the relationship between syntactic complexity of L2 writing and writing quality as judged by human raters, as well as the role of topic in the relationship. It was found that topic had a significant effect on syntactic complexity features of the essays, with one topic eliciting a higher amount of subordination (finite and non-finite) and greater global sentence complexity and the other eliciting more elaboration at the finite clause level (in particular, coordinate phrases and complex noun phrases). Moreover, the less prominent local-level complexity features for essays on one topic tended to have a stronger correlation with scores for that topic. In line with previous research, Crossley et al. (2016) gathered written essays by first-year university students across a semester-long upper-level English for Academic Purposes (EAP) course. Data analysis showed that from the beginning until the end of the semester, participants in this study generally wrote essays that demonstrated greater local, global and text cohesion. The results of the study were treated as an indication of the

fact that indices of cohesion are predictors of human judgments of text organization and overall essay quality for L2 writing. However, cohesion patterns between the longitudinal analysis and the human judgments of quality shared few similarities, indicating a potential mismatch between cohesion growth and assessments of proficiency. It is worth highlighting at this point that most of the existing studies on the influence of lexical features on L2 writing development either report exclusively on university student scripts or fall short in providing evidence of validated rater judgments.

In research on second-language writing assessment, several researchers have examined the textual characteristics of adult student essays across different types of writing tasks (Crossley and McNamara 2011, 2012; Crossley et al. 2011; Grant and Ginther 2000; Guo et al. 2013; Yu 2009) showing that lexical features could predict L2 writing quality assessed by expert raters and automated scoring tools in high-stakes assessment contexts. For example, Grant and Ginther (2000) analysed a specific set of lexical features in 90 EFL learner essays written at three proficiency levels of the TOEFL Test of Written English (TWE). Data analysis revealed differences among the three proficiency levels; as proficiency increased, there was a steady increase in lexical specificity (type-token ratio, average word length), conjuncts, emphatics and amplifiers. In another study, Cumming et al. (2005) examined the textual characteristics of students' essays on the Test of English as a Foreign Language (TOEFL) and observed differences in the textual characteristics of student essays across different types of writing tasks. Differences were also observed in the textual characteristics of adult student essays across judged achievement levels, with higher-scored ones being generally longer and including more clauses and summaries. More recently, Knoch et al. (2014) also examined the textual characteristics within TOEFL iBT adult student essays related to types of writing tasks and judged proficiency levels in order to verify whether the discourse produced in response to the independent and integrated writing tasks differs and further identifies features of written discourse that are typical of different scoring levels. Similar to Cumming et al. (2005), Knoch et al. (2014) found differences in the textual characteristics of adult EFL students' essays across writing tasks and also observed different patterns of textual characteristics across different types of tasks as well as across different achievement levels. Similar results were reported by Banerjee et al. (2007) who explored textual characteristics of adult student essays across different writing tasks and different levels of judged achievement on the International English Language Testing System (IELTS) and also observed differences in students' vocabulary diversity across different writing tasks. The authors found differences in students' use of grammatical markers (e.g., demonstratives such as "this" or "these") across levels of judged achievement, where students' use of grammatical markers tended to decrease as their judged proficiency level increased, suggesting that other cohesive ties might have come into use. More recently, Guo et al. (2013) explored whether linguistic features, such as lexical sophistication, syntactic complexity, cohesion and basic text information, could predict second language writing proficiency in the Test of English as a Foreign Language (TOEFL iBT) integrated and independent writing tasks. The results of this study showed that, amongst other features, lexical sophistication could be used to significantly predict essay scores in the integrated as well as the independent writing tasks, whereas at the same time, given the similarities and differences in the two sets of predictive features, evaluation of the two writing tasks could rely both on similar and distinct features.

Despite the existing literature on the relationships between textual characteristics of student essays and levels of judged writing proficiency, research that has explored the relationships between essay features and indicators of the quality of ratings in younger EFL learners' written texts (aged 13 to 15 years) produced under standard exam conditions is, to the best of our knowledge, either non-existent or not publicly available. More specific information about the degree to which various essay features are associated with rating quality is needed in order to allow young students' levels of English language written competence to be measured effectively and efficiently over time, opening the door to future learning in a broad range of subjects since analysing foreign language writing development

is important in elementary and secondary school children for successful interventions to take place whenever needed (Haswell 2000; McNamara et al. 2010b). At the same time, echoing the views of Cumming et al. (2005) and Knoch et al. (2014) that patterns found in the writing of young students at different levels can be regarded as applicable to their L2 writing development as they advance in proficiency, this study, motivated by the absence of conclusive research data on the writing development of young EFL learners, was carried out in order to assess their writing competence as a function of their performance in the APTIS for Teens Task 4 writing task and further explore the relationship, if any, between textual characteristics and rater-mediated writing assessment.

2. Materials and Methods

In order to investigate linguistic and discourse features present at the different language competence levels of APTIS for TEENS written scripts, the following research questions formed part of the present study:

1. Are there any statistically significant linguistic and discourse differences in essays written by Lower Secondary EFL learners at different APTIS for TEENS performance levels?
2. Which text variables can better predict text complexity variation between high- and low-achievers in the APTIS for TEENS writing test?

The APTIS for TEENS Learners' Written Corpus: The aim of the present study was to investigate L2 learners' linguistic and discourse knowledge based on their written production in the APTIS for TEENS writing test. APTIS for TEENS is a relatively new English assessment exam developed by the British Council that has been designed specifically for teenagers and aims at testing their skills through familiar topics and scenarios. Targeted at 13–17 year-olds, APTIS for TEENS is, according to its creators, a modular and flexible test that enables international EFL learners and educators alike to target specific skills and receive fast results. Since questions have been designed to reflect activities that occur in a teenager's everyday life, such as social media, homework and sports, tasks are expected to be more familiar to test-takers in order for them to focus purely on their English skills and talk, write, speak and listen with ease. According to its designers, this modern approach to testing allows candidates to fully demonstrate their English knowledge and skills. Having said that, APTIS for TEENS can be integrated into current education systems, allowing students' levels of English language competence to be measured effectively and efficiently over time, while opening the door to future learning in a broad range of subjects. APTIS for TEENS can be used by lower and upper secondary schools, ministries of education around the world, language schools, private tutors or teachers and bi-lingual schools in order to streamline learners according to proficiency level within specific language learning programmes, assess their readiness for taking high-stakes certified exams or participating in study abroad and homestay programmes, evaluate learning programmes and test students' progress over extended periods of study. All in all, in the present study, a stratified random sample of 800 APTIS for TEENS written scripts under standard exam conditions on Task 4 were analyzed with 150 scripts per level of performance, i.e., B1.1, B1.2, B2.1. B2.2 and 100 scripts for C1 and C2 levels (see Appendix A for a complete list of text variables). Our decision to stratify the sample in terms of *text length* was necessary in order to avoid text length effects due to size variations, i.e., only texts with 250 words ($\pm$10%) were included in the present corpus. Finally, the writing prompt asked all pertinent candidates to write an argumentative essay of 220–250 words in response to the following statement, i.e., "International sports competitions such as the Olympics help to bring countries together".

Automated Text Analysis Tools: Landmark advances in computational linguistics and machine learning systems have made it possible to go beyond surface text components and adopt more theoretically sound approaches to text complexity, focusing on a wider range of deep text features that take into account semantic interpretation and the construction of mental models and can thus offer a principled means for test providers and test-takers alike to assess this aspect of test construct validity (Graesser et al. 2004, p. 193). In the present

study, a range of available computer programs, such as Coh-Metrix 3.0, VocabProfile 3.0, Computerized Propositional Idea Density Rater 3.0 (CPIDR) and Stylometrics, have been used to automatically measure a variety of text characteristics and provide evidence of distinguishing between written texts assigned to different levels of language proficiency within the context of the APTIS for TEENS exam. To begin with, a great extent of the present study has been based on in-depth textual analysis provided by Coh-Metrix 3.0, a freely available web-based tool developed at the Institute of Intelligent Systems of the University of Memphis, that uses lexicons, parts-of-speech classifiers and statistical representations of world knowledge to measure cohesion and text difficulty at deeper language levels (Crossley et al. 2007, p. 19; Crossley et al. 2011, p. 562; Graesser et al. 2004, p. 194; McNamara et al. 2010a, p. 293). The original goal of its designers was to enhance reading comprehension in L1 classrooms by providing a means to improve textbook writing and to better match textbooks to the intended students (Graesser et al. 2004, p. 194; Louwerse et al. 2004, p. 844; McNamara et al. 2011, p. 380). Nevertheless, more recently, the applicability of the specific software has been extended to cover aspects of foreign language learning. According to its creators, Coh-Metrix is an improvement over conventional readability measures, because it succeeds in examining deeper linguistic text features, such as semantic relatedness and word sense associations, and eventually matches this textual information to the background knowledge of the reader (McNamara et al. 2011, p. 380). As Crossley et al. (2008) repeatedly argued, Coh-Metrix is well-suited to address many of the criticisms of traditional readability formulas because the language metrics it reports on include text-based processes and cohesion features that are integral to cognitive processes, such as syntactic parsing and meaning construction (Perfetti 1992, p. 146; Rayner et al. 2011, p. 116).

VocabProfile 3.0 and its updated BNC version 3.2, two freeware available web-based vocabulary profile programs, were also used in order to estimate word frequency and obtain lists of word tokens (total number of running words), word types (different word forms) and word families (groups containing different forms of a word) for each text in the corpus (Cobb 2007, p. 38, 2010, p. 182). Based on Nation's frequency lists, VocabProfile classifies the vocabulary of a text into frequency levels and outputs a profile that describes the lexical content of the text in terms of frequency bands by showing how much coverage of the text each of the twenty lists accounts for (Cobb 2010, p. 181; Meara 2005, p. 32; Nation 2006, p. 59; 2001, pp. 34–35). In simple terms, 20 ready-made lists are nowadays available with the first including the most frequent 1000 word families of English, the second including the second 1000 most frequent word families and so on (Cobb 2007, p. 38). All of these lists include the base and derived forms of words, so the first 1000 words consist of approximately 4000 forms (Nation 2001, p. 35). The assumption that lies behind the idea of word-families is that an EFL user, who knows well at least one of the members of a family, is more likely to understand other family members, by using knowledge of the most common and regular of the English word building devices (Nation 2006, p. 67). The first three of Nation's lists (i.e., the 3000 most frequent word families) represent the current best estimate of the basic lexicon of EFL users (Cobb 2007, p. 41). VocabProfiler has been used in a variety of studies and has proved particularly useful in helping teachers and students alike prioritize vocabulary worth spending time on teaching and learning in various language courses (Cobb 2007, p. 44; Laufer and Nation 1995, p. 312; Nation and Wang 1999, pp. 358–59; Webb 2010, p. 505). In fact, as Cobb (2010, p. 181) explained, the VocabProfile computer program could help EFL users develop their lexical knowledge at a particular level rather than randomly. It also makes it possible for teachers to devise plausible sequences of lexical acquisition by targeting specific lexical needs and even modifying the learning burden specific texts might place on their students (ibid: p. 181).

Propositional idea density was estimated using the Computerized Propositional Idea Density Rater 3.0 (CPIDR), a computer program that determines the propositional idea density (P-density) of an English text on the basis of part-of-speech tags (Brown et al. 2008, p. 540). Developed at the Institute for Artificial Intelligence of the University of Georgia, the Computerized Propositional Idea Density Rater is a user-friendly Windows

application distributed as open-source freeware through http://www.ai.uga.edu/caspr, accessed on 20 July 2022. To the best of our knowledge, it is the only software that makes such a complex measurement possible. Following Kintsch's theory of comprehension and the representation of meaning in memory (Kintsch 1988, p. 165), CPIDR 3.0 functions based on the idea that propositions correspond roughly to verbs, adjectives, adverbs, prepositions and conjunctions. Thus, after tagging the parts of speech using MontyLingua (Liu 2004, p. 1), CPIDR applies numerous rules following Turner and Greene's handbook of propositional analysis (1977, p. 2) and provides a propositional idea density score by dividing the number of propositions to the total number of words in a text. Tested against human raters' propositional analysis of 80 samples of spontaneous speech, CPIDR was found to agree with the consensus of 2 trained human raters better than the team of 5 trained raters agreed with each other (Brown et al. 2008, p. 543; Covington 2007, p. 6). In fact, as its creators highlighted, by automatically estimating propositional density, CPIDR can open up the possibility of developing more reader-friendly documents and more standardized reading assessments (Brown et al. 2008, p. 544). Following their suggestion, CPIDR was used in the present research as a means of identifying propositions across EFL learners' written texts in an objective and consistent way, since human measurement would inevitably have been subject to personal variation.

In order to assess lexical diversity, Malvern and Richards' (1997, p. 59) D-formula incorporated into the *vocd* command of the Computerized Language Analysis (CLAN) suite of programs of the Child Language Data Exchange System (CHILDES) project was used (MacWhinney 2000, p. 110; MacWhinney and Snow 1990, p. 458; Malvern and Richards 2002, p. 90). A minimum sample size of 50 words is required for *vocd* to compute a valid D, a measurement based on an analysis of the probability of new vocabulary being introduced into longer and longer samples of speech or writing. First, data had to be transcribed in a standard format (CHAT) following a specific coding system, but once the text had been checked and the coding accepted, the software found the best fit between the theoretical model and the empirical data by following a curve-fitting procedure that adjusted the value of the parameter D in the equation until a match was obtained between the actual curve for each passage and the closest member of the family of curves represented by the mathematical model. After calculating the average of Ds three times, *vocd* reported a final optimum D value for each text (Malvern et al. 2004, pp. 63–75; Richards and Malvern 2007, p. 80). This value of the parameter for best fit is the index of lexical diversity, with high values of D reflecting a higher level of lexical diversity and thus a richer vocabulary, whereas word repetition produces lower values of D (McCarthy and Jarvis 2010). Texts with a high D value are expected to be comparatively more difficult to produce because many unique words need to be encoded and integrated with the discourse context. On the other hand, low diversity scores indicate that words are repeated many times in a text, which should generally increase the ease and speed of text processing (MacWhinney 2000, pp. 110–13; Malvern and Richards 2002, p. 90; Malvern et al. 2004, p. 56; McKee et al. 2000, p. 326).

Finally, Stylometrics, a powerful PERL scripted text-analysis tool developed by Prof. Mikros at the National and Kapodistrian University of Athens (Mikros 2011, pp. 130–31), was employed in order to estimate a number of additional text variables, such as word length frequency spectrum for units containing 1 to 14 letters, percentage of apax and dis legomena and relative text entropy. Stylometrics has already been used in a variety of studies on automatic authorship identification (Markopoulos et al. 2011; Mikros 2011, p. 131; Mikros 2009, p. 63; Mikros 2007, p. 463; Mikros and Perifanos 2011, p. 2) and has proved particularly useful in providing text metrics not available in any other commercial software.

With specific reference to the purposes of the present research, categorizing the lexicon of APTIS for TEENS written texts on the basis of frequency bands and describing their linguistic profile per level of language competence can be particularly useful for script raters to consistently assign scores to texts taking into account specific criteria of word frequency. Most importantly, a lexical profile per level of competence could prove beneficial for EFL

teachers and learners alike regarding the depth and breadth of vocabulary knowledge learners need to acquire in order to produce texts in those lexical domains that need to be covered when sitting for a specific exam.

Selection of Text Variables: The 90 text variables identified for investigation in the present research were chosen for both practical and theoretical reasons. First, from the practical standpoint of comparability, it was important to establish whether particular features existed, whose presence in the APTIS for TEENS written exam might introduce constructive relevant variance into assigned scores. At the same time, from a theoretical perspective in the present study, the investigation of the textual characteristics of students' essays was based on a view of writing as a reflection of two main characteristics: linguistic knowledge and discourse knowledge (Grabe and Kaplan 1996; Weigle 2002). As already stated in the literature review section, linguistic knowledge reflects commonly accepted features of language, such as grammar, frequency, concreteness and syntax. To be more specific, grammatical sentence complexity is frequently measured through word and sentence length, subordination, coordination and other structures, such as verb phrases and nominals (Bardovi-Harlig and Bofman 1989; Ferris 1994; Kobayashi and Rinnert 1992; Kuiken and Vedder 2017; Lu 2011). On the other hand, discourse knowledge emphasizes expected conventions within a specific discourse style (i.e., vocabulary and phrases for expressing arguments and making references), and the level of maturity in displaying discourse knowledge is often characterized by how the text is organized (Scardamalia and Paris 1985). It is worth highlighting at this point that, given the inherent complexity of the writing process, it was imperative to include a comprehensive list of text features in order to minimize the risk of omitting variables that might have contributed to attested written performance. Given that previous research has failed to produce a definite set of text variables, no decision was a priori made in terms of their expected significance; i.e., all text elements that relate to linguistic and discourse knowledge and can be measured through available automatic tools were included in the analysis.

3. Results

The present section reports on the text analysis results and looks into text complexity by reporting on the differences across B1, B2, C1 and C2 APTIS for TEENS written texts with regard to specific linguistic features (see Appendix A for the complete list of text variables). Regarding the statistical procedures employed in the present research, it is worth mentioning that, in order to avoid contamination of results due to text length variation, a word limit of 250 words/text (±10%) was followed. In addition, all percentages are reported as valid percentages with missing data excluded. Basic descriptive information is provided through the mean, median and standard deviation estimates, which indicate average values and variability for each data set. Finally, IBM SPSS 28.0 statistical package data were used to compute descriptive statistics and perform Pearson product moment correlations and independent sample *t*-tests.

3.1. Basic Text Information

To obtain a rough idea of surface text features present in the APTIS for TEENS essays, a preliminary analysis of frequencies at the word and sentence level was performed per level of competence, i.e., B1, B2, C1 and C2. To begin with, descriptive statistics showed that the 300 B1 scripts contained an average number of 239 words (SD = 11.51), 9 sentences (SD = 3.75) and 66.15% of different word types (SD = 0.076). They included simpler, familiar syntactic structures (Mean: 15.61, SD = 18.81) and concrete words (Mean: 53.34, SD = 26.95), while their referential cohesion was high in ideas (Mean: 82.55, SD = 20.44) but low in connectivity (Mean: 4.91, SD = 11.17), an indication that the same ideas kept overlapping across sentences without the explicit presence of adversative, additive and comparative connectives to express such relations in the text. Moreover, the specific texts were characterized by the presence of very short words[1], with a mean value of 4.34 letters (SD = 0.26) and 1.42 syllables per word (SD = 0.08). Regarding basic sentence structure

features, B1 texts contained an average of 32 words per sentence (SD = 22.17) but had large variation in terms of the length of sentences within each text with great interchange between some very short and some very long sentences (Standard Deviation of the Mean Length of sentences DESSLd = 17.38).

Regarding the 300 B2 scripts, data analysis showed that texts produced on the same task contained an average number of 241 words (SD = 10.01), 9 sentences (SD = 3.27) and 67.37% of different word types (SD = 0.064). They included simpler, familiar syntactic structures (Mean: 15.77, SD = 19.61) and concrete words (Mean: 50.27, SD = 26.55), while their referential cohesion was high in ideas (Mean: 78.28, SD = 22.66) but low in connectivity (Mean: 3.90, SD = 9.94), an indication that the same ideas kept overlapping across sentences without the explicit presence of adversative, additive and comparative connectives to express such relations in the text. Nevertheless, B2 texts had lower referential cohesion than their B1 counterparts, which could be an indication of reduced repetition of similar ideas. Moreover, the specific texts were characterized by the presence of short words, with a mean value of 4.50 letters (SD = 0.24) and 1.46 syllables per word (SD = 0.07). Regarding basic sentence structure features, B2 texts contained an average of 30 words per sentence (SD = 18.55) and had lower variation in terms of the length of sentences within each text with lower interchange between some very short and some very long sentences (Standard Deviation of the Mean Length of sentences = 14.78).

In contrast to B2 level texts, data analysis showed that texts produced at the C1 level included slightly more complex syntactic structures (Mean: 17.67, SD = 19.36), and their referential cohesion was lower in repetition of the same ideas (Mean: 62.92, SD = 25.62) and higher in connectivity (Mean: 5.20, SD = 7.50), an indication that different ideas kept appearing across sentences with the explicit presence of adversative, additive and comparative connectives to express such relations in the text. Moreover, C1 texts contained an average number of 246 words (SD = 14.79), 11 sentences (SD = 4.55) and 70.53% of different word types (SD = 0.058). In addition, the specific texts were characterized by the presence of short words, with a mean value of 5.50 letters (SD = 0.25) and 1.51 syllables per word (SD = 0.08). Regarding basic sentence structure features, C1 texts contained an average of 28 words per sentence (SD = 13.87) and had lower variation in terms of the length of sentences within each text with lower interchange between some very short and some very long sentences (Standard Deviation of the Mean Length of sentences = 12.41, SD: 6.42).

In accord with the C1 level texts, data analysis showed that texts produced at the C2 level included slightly more complex syntactic structures (Mean: 17.18, SD = 12.09), and their referential cohesion was even lower in ideas (Mean: 59.11, SD = 38.18), an indication that different ideas kept appearing across sentences. At the same time, the very low percentage in connectivity (Mean: 0.58, SD = 0.55) shows that logical relations in the specific texts were displayed in a subtler manner without the explicit presence of adversative, additive and comparative connectives to express such relations in the text. Moreover, C2 texts contained an average number of 244 words (SD = 11.16), 10 sentences (SD = 3.96) and 73.04% of different word types (SD = 0.071). In addition, the specific texts were characterized by the presence of a bit longer words, with a mean value of 6.72 letters (SD = 0.26) and 1.61 syllables per word (SD = 0.07). Regarding basic sentence structure features, C2 texts contained an average of 26 words per sentence (SD = 7.78) and had lower variation in terms of the length of sentences within each text with lower interchange between some very short and some very long sentences (Standard Deviation of the Mean Length of sentences = 11.89, SD: 3.31).

In order to explore and further determine the statistical significance of these findings, independent sample t-tests were carried out and significant differences across the three sets of groups, i.e., B1 $\neq$ B2, B2 $\neq$ C1, C1 $\neq$ C2, were found regarding both "superficial" text features and text easability principal components that go beyond traditional readability measures by providing metrics of text characteristics on multiple levels of language and discourse. To begin with, B1 and B2 texts were found to be statistically different in three

"superficial" text variables, that is, mean number of syllables in words, mean number of letters in words and mean number of words in sentences. More specifically, texts used at the B2 level included a significantly lower number of sentences (t = 2.690, df = 598, $p = 0.007$) than their B1 counterparts and were also characterized by significantly longer words in terms of mean number of syllables (t = 5.093, df = 598, $p < 0.001$) and mean number of letters in words (t = 7.485, df = 598, $p < 0.001$) as well as significantly lower standard deviation of the mean number of syllables in words; i.e., there was less variation in terms of short and long words in these texts (t = −2.187, df = 598, $p = 0.029$). Most importantly, based on the Text Easability Principal Components (Graesser et al. 2011), B1 texts were found to have a significantly higher percentage of referential cohesion, i.e., words and ideas that kept overlapping across sentences and the entire text (t = 2.243, df = 598, $p = 016$), more concrete content words (t = 2.823, df = 598, $p = 0.005$) and a higher percentage of overlapping verbs in the same text (t = 2.096, df = 598, $p = 0.036$).

When comparing B2 with C1 texts, significant differences were found regarding the higher percentage of longer words in terms of syllables per word (t = −3.396, df = 64, $p < 0.001$) and letters per word (t = −3.894, df = 64, $p < 0.001$) included in advanced scripts opposed to the higher percentage of narrative features, i.e., familiar words and common expressions used in oral, everyday language, that characterized B2 texts (t = 4.328, df = 64, $p < 0.001$) along with their significantly higher percentage of overlapping verbs in the text (t = 2.970, df = 64, $p = 0.004$). This could indicate that C1 texts tended to be more complex with increased causal mechanisms, while the B2 texts were likely to contain a higher number of simple forms of narrative with a clear time connection between events.

In accord with our expectations, such a preliminary analysis, albeit useful, was per se limited since such traditional text metrics of word and sentence length are considered rather "superficial" in nature and can provide us with limited information regarding the effect more in-depth textual features can have on overall text complexity. To this end, as shown in the sections that follow, a range of more in-depth text features were analyzed in order to detect more profound linguistic differences across the different levels of English language competence.

3.2. Word Frequency Analysis

Word frequency is an important measure of text complexity as there is increasing research evidence that high-frequency words are normally read more quickly and are more easily produced than infrequent ones. In the present research, a word frequency profile was created for all texts contained in our corpus, and the scores for average frequency for content words, average frequency for all words and average minimum word frequency in sentences obtained through Coh-Metrix, were 2.60 (SD = 0.14), 3.25 (SD = 0.10) and 1.00 (SD = 0.79), respectively, which shows that mainly frequent words were present in B1 texts. Similar results were obtained for B2 texts in which average frequency for content words, average frequency for all words and average minimum word frequency in sentences were 2.52 (SD = 0.13), 3.21 (SD = 0.09) and 1.03 (SD = 0.72), respectively, along with C1 texts in which average frequency for content words, average frequency for all words and average minimum word frequency in sentences were 2.48 (SD = 0.10), 3.12 (SD = 0.07) and 1.52 (SD = 0.70). A slight difference was noted for C2 texts in which the average word frequency for content words was slightly lower at 2.48 (SD = 0.06) along with average minimum word frequency in sentences (1.13, SD = 0.70).

The statistical significance of these contrasts was further explored through independent samples t-tests, and significant differences were found. To be more specific, B1 texts included a significantly higher proportion of frequent words present in the CELEX database (t = 5.360, df = 598, $p < 0.001$) than their B2 counterparts, whereas B2 scripts included a significantly higher percentage of frequent words (t = 4.125, df = 64, $p = 0.001$), including frequency of content words (t = 3.246, df = 64, $p = 0.002$), when compared with C1 texts. No statistically significant differences at the 0.05 level were found between C1 and C2 level texts that could be partly attributed to the very limited number of such texts produced

by learners aged 13–15 years old. These findings show that word frequency can be an important indicator of text complexity that might be of practical usefulness to script raters, since more advanced texts are expected to be characterized by the progressively higher presence of less frequently used words.

3.3. Lexical Richness Analysis

The lexical richness of B2 and C1 level texts was measured through four indices, i.e., percentage of types and tokens for all words and content word lemmas along with lexical diversity optimum LDVOCD and MTLD values. The across levels analysis levels showed that C2 texts contained a more diverse vocabulary since they were characterized by a higher percentage of unique words (Mean = 54.23, SD = 0.049) than their C1 counterparts (Mean = 52.74, SD = 0.039) along with a higher type-token ratio of content words (Mean = 73.04, SD = 0.071) and higher MTLD (Mean = 71.95, SD = 14.13) and VOCD (Mean = 83.08, SD = 14.87) values. On the other hand, all lexical diversity values were lower for B1 texts, which could be treated as an indication of the less diverse vocabulary present in the specific discourse (Richards and Malvern 2007).

Taking the analysis a step forward, the statistical significance of these differences was further investigated by carrying out a set of independent samples t-tests. To be more specific, B2 texts contained a significantly higher proportion of different types of words ($t = -5.131$, $df = 598$, $p < 0.001$) than their B1 counterparts, a finding that was further confirmed by the significantly higher score of MTLD ($t = -10.564$, $df = 598$, $p < 0.001$) and LDVOCD ($t = -10.110$, $df = 598$, $p < 0.001$) indices. Similarly, C1 texts contained a significantly higher proportion of different types of words than their B2 counterparts, a finding that was confirmed by the significantly higher score of MTLD ($t = 3.601$, $df = 64$, $p < 0.001$) and LDVOCD ($t = 3.420$, $df = 64$, $p < 0.001$) indices. Statistically significant differences at the 0.05 level were not found between C1 and C2 level texts that could be attributed to the very limited number of such texts produced by learners aged 12–15 years old when sitting for the APTIS for TEENS exam. Nevertheless, the rest of the findings are considered particularly useful in providing a valid way of discriminating between texts of varying levels of lexical richness and might prove helpful to script raters when facing difficulties in assigning a score due to such text subtleties.

3.4. Text Abstractness Analysis

The extent of abstractness in APTIS for TEENS written texts was measured through eight indices, i.e., age of acquisition for content words, familiarity and concreteness for content words, imageability for content words, meaningfulness for content words, polysemy for content words, mean verb and noun hypernym values. The statistical significance of differences was investigated by carrying out a set of independent samples t-tests. To be more specific, B2 texts contained a significantly higher proportion of less familiar words ($t = 5.579$, $df = 598$, $p < 0.001$) than their B1 counterparts, a finding that was further supported by the significantly lower percentage of concrete words ($t = 2.883$, $df = 598$, $p < 0.001$) with low imageability, i.e., more abstract words for which learners find it difficult to construct a mental image ($t = 4.611$, $df = 598$, $p < 0.001$). Moreover, B2 texts included words with higher age-of-acquisition scores ($t = -9.240$, $df = 598$, $p < 0.001$) along with a higher percentage of polysemous ($t = 3.303$, $df = 598$, $p < 0.001$) and lower levels of hypernymy for nouns ($t = 5.941$, $df = 598$, $p < 0.001$) and verbs ($t = 3.932$, $df = 598$, $p < 0.001$). Words with higher age-of-acquisition scores denote words that are learned later by children, whereas word polysemy is considered to be indicative of text ambiguity because the more senses a word contains, the higher the potential for a greater number of lexical interpretations (for example the word *bank* has at least two senses, one referring to a building or institution for depositing money and the other referring to the side of a river). Finally, lower levels of hypernymy for nouns and verbs in B2 texts reflect an overall use of less specific words in comparison with B1 texts. To be more specific, a word with more hypernym levels invokes more word associations and is thus more concrete (for example the noun *chair* could be

associated with various concepts such as *object*, *furniture*, *seat*, etc.) and easier for EFL users to acquire, whereas a word with fewer hypernym levels is more abstract and thus expected to be more difficult to recall (Crossley et al. 2009, p. 310; 2010, p. 582; 2011, p. 564; Ellis and Beaton 1993, pp. 565–66; Gee et al. 1999, p. 495; Salsbury et al. 2011, p. 346; Schmitt and Meara 1997, p. 27; Zareva 2007, p. 126).

Quite similar differences were noted between B2 and C1 texts, with the latter containing a statistically significant proportion of words with higher age-of-acquisition scores (t = 2.673, df = 64, $p = 0.010$), less familiar (t = 4.397, df = 64, $p < 0.001$) or concrete ones (t = 2.492, df = 64, $p = 0.016$) with lower levels of hypernymy for nouns (t = 2.637, df = 64, $p < 0.001$) along with a higher percentage of polysemous (t = 2.212, df = 64, $p < 0.001$) and lower levels of hypernymy for nouns (t = 3.268, df = 64, $p < 0.001$) scores. Statistically significant differences at the 0.05 level were not found between C1 and C2 level texts that could be greatly attributed to the very limited number of such texts produced by learners aged 13–15 years old when sitting for the APTIS for TEENS exam. Finally, it needs to be pointed out that the inclusion of more abstract, polysemous words at higher levels of language proficiency could be an indication of increased language competence on behalf of the APTIS for TEENS test-takers (Dufty et al. 2006, p. 1253; Cacciari and Glucksberg 1995, p. 291; Crossley et al. 2009, p. 322; Salsbury et al. 2011, p. 352; Schwanenflugel et al. 1997, p. 545).

3.5. Syntactic Complexity Analysis

Syntactic complexity was investigated through a number of metrics that assess the syntactic composition of sentences and the frequency of particular syntactic classes in each text. To be more specific, one set of analysis included eight text indices that can provide clues about text organization, i.e., percentage of all connectives present in a text with them being further divided into positive and negative additive, temporal, causal, contrastive and logical ones. Data analysis showed that B2 compared to B1 texts contained a significantly higher proportion of contrastive (t = 2.704, df = 598, $p = 0.007$) and temporal (t = 2.730, df = 598, $p = 0.007$) connectives, both of which could have been used to clarify relationships among opposing ideas and provide a clear temporal pathway for the readers to follow (Britton et al. 1982, p. 51; Caron et al. 1988, p. 309; Geva 1992, p. 731; Haberlandt 1982, p. 243; Zadeh 2006, p. 1). Although the independent sample t-test revealed only these two statistically significant differences across all levels of texts, the differences mentioned above may be taken to reflect a general tendency of intermediate EFL learners to indulge in more cognitively demanding processes, given the higher percentage of logical arguments and the frequent shift in time sequence in their scripts. Regarding syntactic structure similarity, data analysis revealed statistically significant differences only between B1 and B2 texts regarding the percentage of words before the main verb (t = 2.134, df = 598, $p = 0.003$) and the number of modifiers per noun phrase (t = 2.938, df = 598, $p = 0.007$), both of which scored significantly higher in B2 texts. Statistically significant differences at the 0.05 level were not found between C1 and C2 level texts that could be greatly attributed to the very limited number of such texts produced by learners aged 13–15 years old when sitting for the APTIS for TEENS exam.

The final set of analysis explored the frequency of six additional syntactic pattern density variables, i.e., percentage of noun phrases, verb phrases, adverbial phrases, prepositional phrases, agentless passive voice forms and negations, whose increase has been reported to have a direct impact on text complexity (Charrow 1988, p. 93; Dufty et al. 2006, p. 1254; Gorin 2005, p. 351; Kaup 2001, p. 960; Kemper 1987, p. 323; Kirschner et al. 1992, p. 546; Nagabhand et al. 1993, p. 900; Silver et al. 1989, p. 170). In order to check the statistical significance of any differences, independent samples t-tests were carried out. The alpha level of 0.05 was corrected to 0.007 for this set using the Holm–Bonferroni adjustment model for multiple tests, whereas homogeneity of group variances per text variable was assessed using Levene's Test for Equality of Variances ($p > 0.05$). The analysis revealed statistically significant differences between B1 and B2 texts for three specific syntactic

measures, i.e., proportion of agentless passive voice forms (t = −3.631, df = 598, $p < 0.001$), adverbial phrases (t = −8.924, df = 598, $p < 0.001$) and prepositional ones (t = −4.735, df = 598, $p < 0.001$). In other words, these three features were found to make a significant contribution on distinguishing B1 from B2 texts according to their syntactic complexity and could even be used to obtain a numerical cut-off point between proficiency levels. Thus, they should be given priority in future investigations on syntactic complexity across various levels of language performance. Statistically significant differences were not found between C1 and C2 level texts that could be greatly attributed to the very limited number of such texts produced by learners aged 13–15 years old when sitting for the APTIS for TEENS exam.

3.6. Reference and Cohesion Analysis

Referential cohesion refers to the overlap in content words between local sentences, or *co-reference*. Co-reference is a linguistic cue that can aid readers in making connections between propositions, clauses and sentences in their textbase understanding (Halliday and Hasan 1976; McNamara and Kintsch 1996). Coh-Metrix measures for referential cohesion vary along two dimensions from local to more global ones. Local cohesion is measured by assessing the overlap between consecutive, adjacent sentences, whereas global cohesion is assessed by measuring the overlap across all of the sentences in a text. In the present study, referential cohesion was measured through ten indices, i.e., anaphor overlap (a pronoun that refers to a pronoun or noun in earlier sentences) between adjacent sentences and across all sentences along with noun, argument, stem and content word overlap between adjacent sentences and noun, argument stem and content word overlap across all sentences.

Statistical analysis revealed a significantly higher percentage of content word overlap between adjacent sentences (t = −3.160, df = 598, $p < 0.001$) and across all sentences in B1 rather than B2 texts which indicates increased repetition of the same content words within a text. In a similar vein, B2 texts contained a significantly higher proportion of argument overlap between adjacent sentences (t = 2.487, df = 64, $p < 0.001$) and across all sentences (t = 2.494, df = 64, $p < 0.001$), along with increased content word overlap between adjacent sentences (t = 2.236, df = 64, $p < 0.001$) and anaphor overlap at a local (t = 2.483, df = 64, $p < 0.001$) and global level (t = 2.367, df = 64, $p < 0.001$). These findings were in agreement with our expectations and evidence from previous research that overlapping of word units is more extensive in lower level texts, whereas a higher density of pronouns is pertinent to more complex, advanced texts (Crossley and McNamara 2009, p. 124; Douglas 1981, p. 101; Field 2004, p. 121; Horning 1987, p. 58; Rashotte and Torgesen 1985, p. 186; Rayner and Juhasz 2004, pp. 350–51).

The degree of causal, temporal, spatial and intentional relations in our APTIS for TEENS corpus was investigated through five relevant indices provided by Coh-Metrix, i.e., incidence of causal and intentional verbs along with causal, intentional and temporal cohesion scores. Although no significant differences were found across the different levels with regard to these cohesion measures, a more in-depth analysis of verb tenses revealed a statistically significant difference for past and present tenses per level of competence. To be more specific, the statistical analysis showed that B1 texts included a significantly higher proportion of present tenses (t = 3.229, df = 598, $p = 0.004$), whereas past tenses were more frequent in B2 texts (t = −2.756, df = 598, $p = 0.008$) and C1 ones (t = −2.103, df = 64, $p < 0.001$). The higher incidence of past tenses in more advanced texts could be interpreted as an indicator of text complexity, as previous research has already shown that such tenses may have a negative impact on the ease and speed of text processing (Nagabhand et al. 1993, p. 900).

Finally, more features of informal, everyday speech were present in B1 rather than B2 texts, such as nouns (t = 3.738, df = 598, $p < 0.001$) and first (t = 6.494, df = 598, $p < 0.001$), second (t = 3.017, df = 598, $p < 0.001$) and third person singular pronouns (t = 2.474, df = 598, $p < 0.001$). On the other hand, B2 texts included a significantly higher percentage of more formal features, such as verbs (t = −3.618, df = 598, $p < 0.001$), gerunds (t = −5.474, df = 598,

$p < 0.001$), first (t = −2.956, df = 598, $p < 0.001$) and third person plural pronouns (t = −5.513, df = 598, $p < 0.001$), than their B1 counterparts, whereas C1 texts included a significantly higher percentage of adjectives than their B2 counterparts (t = −2.245, df = 64, $p < 0.001$).

3.7. Within Levels Text Analysis

An additional set of independent sample t-tests were carried out within each level of language competence, i.e., B1.1 ≠ B1.2, B2.1 ≠ B2.2, and significant differences were found regarding specific text features. More specifically, texts used at the B1.2 level contained significantly longer words in terms of mean number of syllables (t = −2.571, df = 298, $p = 011$) and mean number of letters in words (t = −4.901, df = 298, $p < 0.001$) than their B1 counterparts and were also characterized by a significantly lower percentage of content word overlap (t = −2.905, df = 298, $p = 004$) but higher anaphor overlap, i.e., pronouns in one sentence referred to nouns or pronouns in an earlier sentence thus clarifying connected ideas (t = −2.680, df = 298, $p = 008$). Moreover, B1.1 texts included a significantly higher percentage of conceptually similar sentences (t = 2.634, df = 298, $p < 0.001$), whereas B1.2 texts contained a more diverse vocabulary since they were characterized by a significantly higher percentage of unique words (t = −2.348, df = 298, $p = 020$) along with a higher type-token ratio of content words (t = −3.864, df = 298, $p < 0.001$) and higher MTLD (t = 5.987, df = 298, $p < 0.001$) and VOCD (t = −4.274, df = 298, $p < 0.001$) values. Regarding syntactic pattern density, the analysis revealed statistically significant differences between B1.1 and B1.2 texts for three specific syntactic measures, i.e., proportion of adverbial phrases (t = −4.526, df = 298, $p < 0.001$), prepositions (t = −2.604, df = 298, $p < 0.001$) and infinitives (t = −2.787, df = 298, $p = 0.006$). Finally, B1.1 were rated as much easier by the Flesch Reading Ease index (t = 2.252, df = 298, $p = 0.025$) since they included a significantly higher percentage of nouns (t = 3.287, df = 298, $p < 0.001$), first person (t = 2.978, df = 298, $p < 0.003$) and third person singular pronouns (t = 3.066, df = 298, $p < 0.002$) and more familiar content words (t = 3.416, df = 298, $p < 0.001$), all of which are features of simpler, everyday speech. On the other hand, B1.2 texts contained a significantly higher percentage of adverbs (t = −2.469, df = 298, $p = 0.014$), more third person plural pronouns (t = −5.696, df = 298, $p < 0.001$), less familiar words (t = −5.341, df = 298, $p < 0.001$) and less specific nouns with lower hypernym values (t = −2.520, df = 298, $p = 0.012$).

When comparing B2.1 and B2.2 texts, data analysis showed that B2.2 level texts contained significantly longer words in terms of mean number of syllables (t = −3.505, df = 298, $p = 011$) and mean number of letters in words (t = −4.670, df = 298, $p < 0.001$), whereas B2.1 texts were characterized by a significantly higher percentage of referential cohesion with similar words and ideas overlapping across sentences and the entire text (t = 2.289, df = 298, $p = 0.023$) along with higher anaphor overlap (t = 2.315, df = 298, $p = 0.021$) with content word overlapping between adjacent sentences (t = 3.116, df = 298, $p = 0.002$) and across all sentences in a text (t = 2.689, df = 298, $p = 0.008$). This overlap of words and ideas in B2.1 texts was confirmed by Latent Semantic Analysis metrics which also showed increased semantic overlap between adjacent sentences (t = 2.852, df = 298, $p = 0.005$) and across all sentences in a paragraph (t = 2.447, df = 298, $p = 0.015$). On the other hand, B2.2 texts contained a more diverse vocabulary since they were characterized by a significantly higher type-token ratio of all words (t = −3.476, df = 298, r = 0.27, $p = 0.001$) and higher MTLD (t = −3.904, df = 298, r = 0.27, $p < 0.001$) and VOCD (t = −3.931, df = 298, r = 0.27, $p < 0.001$) values. Regarding syntactic pattern density, the analysis revealed statistically significant differences between B2.1 and B2.2 texts for two specific syntactic measures, i.e., proportion of adverbial phrases (t = −2.157, df = 298, r = 0.36, $p = 0.032$) and gerunds (t = −2.782, df = 298, r = 0.36, $p = 0.006$). Finally, B2.1 were rated as much easier by the Flesch-Kincaid Reading Ease index (t = 2.750, df = 298, $p = 0.006$) and the Second Language Readability index (t = 2.894, df = 298, $p = 0.004$) since they included a significantly higher percentage of first person (t = 2.255, df = 298, $p = 0.025$) and third person singular pronouns (t = 2.976, df = 298, $p = 0.003$) along with more familiar words (t = 3.120, df = 298, r = 0.36, $p = 0.002$) including content words (t = 2.226, df = 298, $p = 0.027$), all of

which are features of simpler, everyday speech. On the other hand, B1.2 texts contained a significantly higher percentage of less familiar words (t = −3.312, df = 298, $p < 0.001$) and words with higher age-of-acquisition scores that denote words that are learned later by children (t = 4.199, df = 298, r = 0.36, $p < 0.001$).

An additional set of analyses was carried out in order to explore the finer differences between B1.2 and B2.1 texts. Independent sample t-test results showed that B2.1 texts contained a significantly higher proportion of diverse vocabulary since they were characterized by significantly higher MTLD (t = −2.695, df = 298, p.007) and VOCD (t = −3.197, df = 298, r = 0.27, p.002) values, sentences were more syntactically complex with a higher percentage of words before the main verb (t = −2.420, df = 298, r = 0.27, p.016), included more adverbial phrases (t = −3.024, df = 298, p.003), passive voice instances (t = −2.192, df = 298, p.029) and less concrete words (t = −2.033, df = 298, p.003) along with more polysemous words (t = −3.542, df = 298, $p < 0.001$) and lower verb (t = −2.235, df = 298, $p < 0.001$) and noun hypernymy (t = −2.033, df = 298, $p < 0.001$) which reflect an overall use of less specific words.

In order to check linguistic development from one edge (B1) to the other (C2) of our language competence imaginary line, a set of analyses was carried out in order to spot significant differences between B1 and C2 texts. In accordance with our expectations, C2 texts included longer, less frequent words as shown in the mean number of syllables (t = −3.885, df = 64, $p < 0.001$) and mean number of letters in words (t = −5.670, df = 64, $p < 0.001$) and had less narrative features (t = 2.535, df = 64, p.014). Moreover, C2 texts had lower referential cohesion, i.e., less overlapping of words and ideas (t = 4.087, df = 64, $p < 0.001$) with a lower percentage of argument overlap between adjacent sentences (t = 2.583, df = 64, p.012) and across all sentences (t = 2.023, df = 64, $p < 0.001$) along with lower content word overlap between adjacent sentences (t = 2.353, df = 64, p.022) and across all sentences (t = 2.099, df = 64, p.040). In addition, C2 texts included a significantly higher proportion of more diverse vocabulary since they were characterized by a significantly higher percentage of unique words (t = −4.706, df = 64, $p < 0.001$), along with higher MTLD (t = −7.665, df = 64, $p < 0.001$) and VOCD (t = −7.795, df = 64, $p < 0.001$) values. Apart from lexical diversity, C2 texts were also more syntactically diverse regarding adjacent sentences (t = −2.058, df = 64, $p = 0.004$) and across all sentences (t = −2.082, df = 64, $p < 0.002$). Finally, C2 texts were characterized by their higher percentage of adverbs (t = −3.218, df = 64, $p = 0.002$), prepositions (t = −3.821, df = 64, $p < 0.001$), passive voice instances (t = −3.560, df = 64, $p < 0.001$), gerunds (t = −4.831, df = 64, $p < 0.001$), verbs (t = −3.748, df = 64, $p < 0.001$) and adjectives (t = −2.338, df = 64, p.023), along with less familiar (t = 5.820, df = 64, $p < 0.001$) and concrete (t = 2.082, df = 64, $p < 0.001$) words with lower hypernymy in both verb (t = −4.529, df = 64, $p < 0.001$) and noun (t = −4.529, df = 64, $p < 0.001$) values.

In a nutshell, the main text features that have repeatedly been found to make a significant contribution on distinguishing texts both within and across levels are word frequency, word abstractness, lexical diversity and lexical and semantic overlap, all of which could be used to obtain a numerical cut-off point between proficiency levels. Thus, they should be given priority in future investigations on writing development across various levels of language performance.

3.8. Text Features and Scores

In order to explore the relationship, if any, between assigned scores (1–6) and text features, Pearson correlation coefficients were estimated. Data analysis showed that assigned scores across the whole set of 800 scripts highly correlated with five specific text variables, that is, word frequency as indicated in mean word length (r = −0.437, $p < 0.001$), semantic overlap across sentences (r = −0.685, $p < 0.001$), lexical diversity (r = 0.548, $p < 0.001$), word familiarity (r = 0.499, $p < 0.001$) and verb and noun hypernymy (r = 0.364, $p < 0.001$). With specific reference to hypernymy, Coh-Metrix uses WordNet to report word hypernymy (i.e., word specificity). In WordNet, each word is located on a hierarchical scale allowing for the measurement of the number of subordinate words below and superordinate words above

the target word. Coh-Metrix provides estimates of hypernymy for nouns (WRDHYPn), verbs (WRDHYPv), and a combination of both nouns and verbs (WRDHYPnv). A lower hypernymy value reflects an overall use of less specific words, while a higher value reflects an overall use of more specific words.

It is worth pointing out that despite the fact that the abovementioned analysis is statistically acceptable, treating all test-takers as one group could have deprived us of valuable information per level of competence, since moderate correlations at one level could have been partialed from non-existing or even negative correlations at another level of language competence. Bearing these limitations in mind, in the following sections, the contribution of text indices to mean writing performance per level of competence is discussed in order to gain a more complete understanding of test-takers' skills within each level. Given the great number of correlations that emerged from the high number of text variables, only medium to high correlations are presented.

Data analysis between B1 and B2 scripts (600 scripts in total/300 per level) showed that higher scores highly correlated with five specific text variables, that is, semantic overlap across sentences ($r = -0.487$, $p < 0.001$), lexical diversity ($r = 0.397$, $p < 0.001$), adverbial phrases ($r = 0.343$, $p < 0.001$), word familiarity ($r = -0.353$ $p < 0.001$) and verb and noun hypernymy ($r = -0.236$, $p < 0.001$). In addition, data analysis between B2 and C1/C2 scripts showed that higher scores highly correlated with six specific text variables, that is, word frequency ($r = -0.458$, $p < 0.001$), text narrativity ($r = -0.476$, $p < 0.001$), referential cohesion ($r = -0.377$, $p < 0.001$), lexical diversity ($r = 0.410$, $p < 0.001$), word familiarity ($r = -0.482$, $p < 0.001$) and verb and noun hypernymy ($r = -0.340$, $p < 0.001$). Moreover, data analysis within the B1 level but between B1.1 and B1.2 scripts showed that higher scores highly correlated with two specific text variables, that is, lexical diversity ($r = 0.328$, $p < 0.001$) and semantic overlap across sentences ($r = -0.801$, $p < 0.001$), whereas within B2 level but between B2.1 and B2.2 scripts, two noteworthy correlations emerged, i.e., word frequency ($r = -0.261$, $p < 0.001$) and lexical diversity ($r = 0.222$, $p < 0.001$). Finally, while trying to trace significant differences between scripts at the B2.2 level and C1 level, it became apparent that higher scores highly correlated with six specific text variables, i.e., text narrativity ($r = -0.364$, $p < 0.001$), word frequency ($r = -0.447$, $p < 0.001$), lexical diversity ($r = 0.391$, $p < 0.001$), word familiarity ($r = -0.488$, $p < 0.001$), verb and noun hypernymy ($r = -0.376$, $p < 0.001$) and word abstractness ($r = 0.321$, $p < 0.001$). The only high correlation that emerged regarding scores between C1 and C2 scripts was semantic overlap across sentences ($r = -0.710$, $p < 0.001$) which could be an indication of ideas repeated within C1 scripts, but EFL writers are able enough to use diverse vocabulary including less frequent and more abstract words.

4. Discussion

The purpose of the present study was to use a range of advanced computational linguistics and automated machine in order to examine to which essays assigned different scores in the APTIS for TEENS exam can be distinguished from one another using a number of text features. The investigation of the textual characteristics of young international EFL students' essays produced under standard exam conditions was based on a view of writing as a reflection of two main characteristics: linguistic knowledge and discourse knowledge (Grabe and Kaplan 1996; Weigle 2002). To be more specific, discourse knowledge related to the presence of cohesive ties created by referencing, conjunction and lexical cohesion. On the other hand, linguistic knowledge explores the occurrence of surface text features such as number of words and sentences per text and word frequency along with lexical diversity and syntactic complexity.

Regarding the first research question (i.e., Are there any statistically significant linguistic and discourse differences between essays written by Lower Secondary EFL learners at different APTIS for Teens performance levels?), data analyses revealed statistically significant differences ($p \leq 0.05$) between the linguistic features present in the essays produced by lower secondary EFL learners at different levels of language performance. Such a finding

could provide initial validity evidence for the specific exam paper and also shed more light on L2 acquisition of lexis. At the same time, regarding the second research question (i.e., *Which text variables can better predict text complexity variation between high- and low-achievers in the APTIS for TEENS writing test?*), the main text features that have repeatedly been found to make a significant contribution to distinguishing texts both within and across levels of English language competence are: word frequency, word abstractness, lexical diversity and lexical and semantic overlap, all of which could be used to obtain a numerical cut-off point between L2 proficiency levels.

In accord with existing research on adult L2 written scripts (Banerjee et al. 2007; Crossley et al. 2011; Crossley and McNamara 2014; Cumming et al. 2005; Grant and Ginther 2000; Knoch et al. 2014; Yang et al. 2015), higher-scored essays in the APTIS for TEENS exam generally included longer and more syntactically complex sentences (i.e., mean length of sentences, number of subordinated clauses, number of coordinated clauses, ratio of clauses per sentence and ratio of subordinated clauses per clause), along with less frequent and more diverse vocabulary. Furthermore, the findings of the present study tally with recent studies that have indicated that local cohesion is negatively related to essay quality such as the one by Crossley and McNamara (2012) who examined the writing quality of independent essays (i.e., impromptu writing) produced by Hong Kong high school students and found that local and text cohesive devices, such as content word overlap between adjacent sentences, positive logical connectives, aspect repetition and semantic similarity between sentences, were negatively correlated with expert ratings of essay quality for this cohort of students. Our findings are also partly in tune with results from Guo et al. (2013), who reported that indices of local cohesion (e.g., content word overlap and conditional connectives) and text cohesion (e.g., aspect repetition) were negatively correlated with judgments of essay quality for the independent essays of the Test of English as a Foreign Language (TOEFL).

More specifically, B1 scripts in the APTIS for TEENS exam included simple, familiar syntactic structures and concrete words, while their referential cohesion was high, an indication that the same ideas kept overlapping across sentences without the explicit presence of adversative, additive and comparative connectives to express such relations. Moreover, the specific texts were characterized by the presence of short words, which is a feature of frequent words as opposed to more sophisticated ones that tend to be longer. As White (2011, p. 87) explained, less familiar words are longer words because they tend to be affixed, context-specific and content-bearing. This comes in agreement with Zipf's law (1935 cited in Carrell 1987, p. 22) that word length is inversely related to word frequency, i.e., longer words tend to be more difficult than shorter ones, especially for L2 learners (Harrison 1999, p. 429; White 2011, p. 87). Moreover, dating back to Thorndike's word frequency thesis, infrequent words are deemed to be less familiar and thus more difficult than more frequent ones (Carrell 1987, pp. 22–23; Haberlandt and Graesser 1985, p. 358). Regarding B2 scripts, data analysis showed that essays produced on the same task included simple, familiar syntactic structures but had lower referential cohesion than their B1 counterparts, which could be an indication of reduced repetition of similar ideas. Referential cohesion refers to overlap in content words between local sentences, or co-reference. Co-reference is a linguistic cue that can aid readers in making connections between propositions, clauses and sentences in their textbase understanding (Halliday and Hasan 1976; Louwerse 2002; McNamara and Kintsch 1996). Coh-Metrix measures for referential cohesion vary along two dimensions. First, the indices vary from local to more global. Local cohesion is measured by assessing the overlap between consecutive, adjacent sentences, whereas global cohesion is assessed by measuring the overlap between all of the sentences in a paragraph or text.

Moreover, the specific texts had lower variation in terms of the length of sentences within each text with lower interchange between some very short and some very long sentences. In contrast with B2 level texts, data analysis showed that C1 texts included slightly more complex syntactic structures, and their referential cohesion was lower in word repetition and higher in connectivity, an indication that different ideas kept appearing across

sentences with the explicit presence of adversative, additive and comparative connectives to express such relations in the text. Finally, C2 texts included slightly more complex syntactic structures, and their referential cohesion was even lower in ideas, an indication that different ideas kept appearing across sentences. A text with high referential cohesion contains words and ideas that overlap across sentences and the entire text, forming explicit threads that connect the text for the reader. Low cohesion text is typically more difficult to produce and process because there are fewer connections that tie the ideas together for the reader. At the same time, the very low percentage in connectivity shows that logical relations in the specific texts were displayed in a subtler manner without the explicit presence of adversative, additive and comparative connectives to express such relations in the text. In addition, the specific texts were characterized by the presence of longer words and had lower variation in terms of the length of sentences within each text with lower interchange between some very short and some very long sentences. At the level of sentence composition, we found a significant increase in sentence coordination and subordination. These findings agree with previous research carried out by Bulté and Housen (2014) on college-level EFL writers in which a significant decrease in simple sentences and a significant increase in compound sentences at the end of an intensive English language programme were reported. At the level of clause linking, there was also a significant increase in both clause coordination and subordination in higher-scored scripts, which is consistent with the results obtained by Lorenzo and Rodríguez (2014) with secondary-level EFL learners, while Bulté and Housen (2014) also found that the number of coordinated clauses per sentence increased significantly in their students' scripts by the end of the programme. Such a finding also agrees with the generally accepted belief that syntactic complexity in language is related to the number, type and depth of embedding in a text. Syntactically simple authors use short, single clause sentences and rely more heavily on coordinated structures to provide cohesion and show relationships. Syntactically complex authors, on the other hand, use longer sentences and more subordinate clauses that reveal more complex structural relationships (Beaman 1984; Lu and Ai 2015). Thus, sentences with complex, embedded clauses can potentially place heavier demands on working memory and are considered more cognitively demanding to create.

Regarding word frequency, B1 texts contained a significantly higher proportion of frequent words than their B2 counterparts. At the same time, B2 scripts included a significantly higher percentage of frequent words, including frequency of content words, when compared with C1 texts. Regarding lexical diversity, the across levels analysis levels showed that C2 texts contained a more diverse vocabulary since they were characterized by a higher percentage of unique words along with a higher type-token ratio of content words. On the other hand, all lexical diversity values were lower for B1 texts, which could be treated as an indication of the less diverse vocabulary present in the specific scripts. Given the fact that lexical diversity refers to the variety of unique words (*types*) that occur in a text in relation to the total number of words (*tokens*), when the number of word types is equal to the total number of words (tokens), then all of the words in a text are different, and its lexical diversity is at a maximum. A high number of different words in a text indicates that new words were integrated into the discourse context. By contrast, lexical diversity estimate is lower when the same set of words is used multiple times across the text.

With reference to word abstractness, B2 texts contained a significantly higher proportion of less familiar words than their B1 counterparts, a finding that was further supported by the significantly lower percentage of concrete words with low imageability, i.e., more abstract words for which learners find it difficult to construct a mental image. Moreover, B2 texts included words with higher age-of-acquisition scores along with a higher percentage of polysemous words and lower levels of hypernymy for nouns and verbs. Words with higher age-of-acquisition scores denote words that are learned later by children, whereas word polysemy is considered to be indicative of text ambiguity because the more senses a word contains, the higher the potential for a greater number of lexical interpretations. Finally, lower levels of hypernymy for nouns and verbs in B2 texts reflect an overall use

of less specific words in comparison with B1 texts. Quite similar differences were noted between B2 and C1 texts with the latter containing a statistically significant ($p \leq 0.05$) proportion of less familiar or concrete words which can be an indication of increased vocabulary competence on behalf of the APTIS for TEENS test-takers.

Regarding syntactic structure similarity, data analysis revealed statistically significant differences ($p \leq 0.05$) between B1 and B2 texts regarding the percentage of words before the main verb and the number of modifiers per noun phrase, both of which scored significantly higher in B2 texts. As far as lexical and semantic overlap is concerned, statistical analysis revealed a significantly higher percentage of content word overlap between adjacent sentences and across all sentences in B1 rather than B2 texts which indicates increased repetition of the same content words within a text. In a similar vein, B2 texts contained a significantly higher proportion of argument overlap between adjacent sentences and across all sentences, along with increased content word overlap between adjacent sentences and anaphor overlap at a local and global level. These findings agree with our expectations and evidence from previous research that overlapping of word units is more extensive in lower level texts, whereas a higher density of pronouns is pertinent to more complex, advanced texts (Crossley and McNamara 2009, p. 124; Douglas 1981, p. 101; Field 2004, p. 121; Horning 1987, p. 58; Rashotte and Torgesen 1985, p. 186; Rayner and Juhasz 2004, pp. 50–51). Furthermore, the higher incidence of past tenses in C1 texts could be interpreted as an indicator of text complexity, as previous research has already shown that such tenses may have a negative impact on the ease and speed of text processing (Nagabhand et al. 1993, p. 900). Finally, features of more informal, everyday speech, such as nouns along with first, second and third person singular pronouns, were present in B1 rather than B2 texts. On the other hand, B2 texts included a significantly higher percentage of more formal features such as verbs, gerunds, first and third person plural pronouns, whereas C1 texts included a significantly higher percentage of adjectives.

To conclude, it might be safe to argue that the results of the present study agree with the findings of the existing literature and further support the notion that there seems to be a close relationship between EFL learners' writing development and their language level (Berninger et al. 1991; Crossley et al. 2011). To be more specific, learning to write begins with the mastery of producing legible letters and basic spelling (Abbott et al. 2010). Once these skills are attained, young EFL writers work to master basic grammar and sentence structure, while during later stages of development, writers begin to focus on text cohesion and syntactic structures (McCutchen 1986; Witte and Faigley 1981) while trying to produce more elaborate texts with more sophisticated words, more complex sentence structure and fewer cohesive features as a function of their L2 lexical growth. Our findings seem to confirm this process of L2 writing development and further support the view of an increasing growth in multiple dimensions of text complexity while uncovering a progress in writing characterised by increased sentence length, lexical sophistication and phrasal elaboration as language proficiency increases. Moreover, the findings of the present study seem to add to our state of knowledge and further support the view that syntactic complexity grows as writers become increasingly more capable of using the L2 with linguistic maturity (e.g., Ai and Lu 2013; Lorenzo and Rodríguez 2014; Lu 2011). Similarly to Mazgutova and Kormos (2015), who investigated the development of the lexical and syntactic complexity of two groups of upper-intermediate EFL students enrolled in a month intensive pre-sessional English for Academic Purposes programme, our findings show that low-performers in the APTIS for TEENS written task fell behind in their employment of phrasal elaboration measures (i.e., noun modification via adjectives and prepositional phrases, complex nominals in subject position, multiple modifiers after the same noun) and in subordination-related measures (syntactic structure similarity, conditionals and relative clauses).

5. Conclusions

In order to contribute to enhancing the scoring validity of the APTIS for TEENS writing test, this study carried out a microanalysis of in-depth text features present in the APTIS for TEENS writing scripts produced by international EFL students under standard exam conditions while exploring their relationship with the assigned scores by specially trained raters. A total of 800 international EFL students aged 13 to 15 years old took part in the official exam, and 800-scored written essays on a specific task prompt of the pertinent test were analysed. The results showed statistically significant differences ($p \leq 0.05$) between the linguistic features identified in the essays produced by young EFL learners at different levels of language competence. The main text features that were repeatedly found to make a significant contribution to distinguishing texts both within and across levels were word frequency, word abstractness, lexical diversity and lexical and semantic overlap, all of which could be used to obtain a numerical cut-off point between proficiency levels. In other words, texts with greater lexical diversity and sophistication along with increased syntactic complexity were assessed more positively concerning their overall quality as well as the analytic criteria of "task fulfilment and topic relevance regarding providing reasons, explanations and supporting detail" along with "essay structure and paragraphing:, "grammatical range and accuracy", "vocabulary range and accuracy" and "cohesion". These findings support the notion that progress in L2 writing is primarily associated with producing more elaborate texts with more sophisticated words, more complex sentence structure and fewer cohesive features as a function of increased language competence. The perceived benefit of such a clearly articulated theoretical and practical position for the writing construct in the APTIS for TEENS test is to provide a rationale for the way in which the British Council operationalises such a construct in the specific tests and essentially to deepen our understanding of the theoretical basis upon which the specific examination system tests different levels of language proficiency across its range of assigned scores.

In terms of research on textual characteristics of student essays, this study contributes to previous research on L2 written development of secondary- and college-level EFL learners (e.g., Ai and Lu 2013; Bulté and Housen 2014; Lorenzo and Rodríguez 2014; Lu 2011; Martinez 2018) and writing assessment in high-stakes exam contexts (e.g., Banerjee et al. 2007; Crossley et al. 2011; Crossley and McNamara 2014; Cumming et al. 2005; Grant and Ginther 2000; Knoch et al. 2014; Yang et al. 2015) by examining associations between these characteristics and rating quality (i.e., assigned scores) but adds to our present state of knowledge by examining performance across different levels of foreign language competence with all participants responding to the same writing prompt and identifying specific text variables that have repeatedly been found to make a significant contribution on distinguishing texts both within and across levels of language proficiency. At the same time, addressing Ortega's (2003) word of caution that one of the problems of many second language studies that have investigated complexity of writing and its relationship with writing quality lies in the fact that they have made use of a small number of scripts, which leads to conflicting findings, the present study made use of a rather large sample of 800 essays produced by international EFL teenagers under standard exam conditions.

Although no advanced alignment between rating quality indicators and textual characteristics of learner scripts was attempted within the framework of Rasch measurement theory, an attempt to empirically identify relationships among these variables and the APTIS for TEENS writing rating scales showed that human raters seemed to assign a low mark (i.e., B1.1, mark:1) when scripts were not fully on topic, writers expressed an opinion and provided one reason but did not provide sufficient explanation or supporting details, paragraph structure was not used appropriately and very simple grammatical structures were present, while frequent errors occurred when attempting complex structures. Moreover, such scripts contained only simple cohesive devises, whereas links between ideas were not clearly indicated, and limitations in vocabulary made it difficult to deal fully with the assigned task. On the other hand, raters assigned a higher mark (B2.1, mark: 3) when

scripts were on topic and learners developed an argument in relation to the topic, providing one reason with sufficient explanation and relevant supporting details. In addition, organization of ideas in the form of paragraphs was mostly appropriately employed and some complex grammar constructions were used accurately, while errors did not lead to serious misunderstanding. These essays were perceived to contain a sufficient range of vocabulary to discuss the topics required by the task, inappropriate lexical choices did not lead to misunderstanding, and a limited number of cohesive devices were used to indicate the links between ideas within and across paragraphs. Finally, in accord with APTIS for TEENS writing rating scales, the raters seemed to assign a very high mark (i.e., C1, mark: 5) when scripts were fully on topic, with clear and well-structured arguments, and writers had managed to highlight and expand on important points while providing clear explanations and relevant supporting details. These essays also contained a range of complex grammatical constructions that were used accurately, while some minor errors occurred but did not impede understanding. Finally, a range of cohesive devices was used to clearly indicate the links between ideas both within and across paragraphs, and a wide range of vocabulary was used to discuss the topics required by the task with slightly inappropriate lexical choices.

To conclude, through its investigation of the significant relationships among a range of text variables, the present research attempted to provide evidence regarding the extent to which essays assigned different scores in the APTIS for TEENS exams can be distinguished from one another, using a number of linguistic features related to both linguistic and discourse knowledge. The study also aspired to make a methodological contribution in that, instead of examining a limited number of text variables independently, it made use of advanced text analysis software applications and investigated the impact of 90 text variables on text complexity, while all participants responded to the same writing prompt under standard exam conditions. Having said that, such an approach partly addressed Bulté and Housen's (2014) concerns that most studies on L2 writing complexity suffer from low content. Thus, treating writing as a highly complex construct consisting of several sub-constructs and components, our study on secondary-level international EFL learners with different levels of language proficiency incorporated measures that captured different aspects of text complexity, including grammatical intricacy, lexical density and syntactic complexity. In addition, the focus on secondary education EFL learners' writing competence in the present study makes a novel contribution as such an age group of international test-takers has received very little attention in the pertinent literature on L2 writing assessment.

A major goal of research on rating quality is to identify areas of rater-mediated performance assessments in need of improvement in order to provide support for the validity of the interpretation and use of ratings (Wind et al. 2017). In the present study, an attempt was made to incorporate the textual characteristics of learner essays into an empirical rating quality analysis that can be used to empirically evaluate the degree to which scores across essays vary based on specific textual characteristics. Having said that, the findings of the study can have several practical implications that are most directly related to rater training and remediation procedures, as well as the development of scoring materials for writing assessment purposes. For example, following the procedures illustrated in this study, organizers of rater training workshops can examine a specific set of text features based on benchmark essays that raters score during their training sessions. Information about these relationships can be used to facilitate discussions about aspects of learners' scripts that might complicate scoring or revise training materials that are better suited to address rating discrepancies or rater misconceptions regarding the appropriate use of a writing marking scale. Relatedly, rater trainers can use empirical evidence of relationships among textual features to identify additional performance-level exemplar essays for training that could include characteristics that raters are consistently struggling to rate. This suggestion is particularly promising for computer-based writing assessments, as no transcription would be required. As another practical application, organizers of rater training workshops could

use the procedures illustrated in this study to examine the range of textual characteristics within a set of student essays in order to identify extreme examples of certain textual characteristics. When combined with information about how these textual characteristics correspond to rating quality, rater trainers can use such "extreme" examples to provide focused training and remediation procedures that reflect the unique characteristics of a particular assessment sample.

As with all studies, the implementation of the present one presented a number of challenges and limitations that we hope will be overcome in future research. For instance, due to the fact that only a specific writing prompt was used, written texts inevitably belonged to a specific genre, i.e., argumentative essay; should the range of prompts increase, the generalizability of the present results might be further strengthened or reduced since different topics may give rise to different text features. It would also be useful to extend the present analysis to scripts produced before and after learners attending specially designed writing programmes for evidence-based conclusions to be drawn following a comparative corpus-based approach. Moreover, analyses that focus on individual raters' judgments (e.g., Wind et al. 2017) could be conducted to examine the alignment between textual characteristics and rater judgment at the level of individual raters. For example, individual raters' judgments could be examined by looking into the correspondence between textual characteristics and fit statistics specific to individual raters. Alternatively, the alignment between individual raters and essay characteristics could be examined using interaction analyses based on the Many-Facet Rasch model (Goodwin 2016). Whereas the current study focused on the overall group of raters' judgments of individual essays as well as the combination of textual characteristics in essay feature profiles, this approach would provide insight into individual rater judgments of the overall group of essays. Additional research that explores the practical utility of incorporating information about relationships between textual characteristics and model-data fit for improving practice in large-scale rater-mediated writing assessment is also needed. These studies, during which ratings on benchmark essays could be used to examine empirical relationships between textual characteristics, are perhaps most feasible in the context of rater sessions. Moreover, future studies could involve additional text characteristics such as spelling and organization along with test-takers' personal characteristics such as gender, years of learning English as a foreign language, medium of instruction and L1 background. It is believed that these studies could improve our understanding of rater judgements of EFL essays and provide empirical data on key aspects of L2 writing assessment

Finally, yet importantly, to the best of our knowledge, neither Coh-Metrix nor any other program currently available includes any indices to measure pragmatic features of sociolinguistic discourse theory. Having said that, the set of text indices used in the current study should be viewed as a limited construct of writing, and the indices included in this analysis should be considered partly representative of the complete range of lexical and discourse features in young EFL learners' essays at an international level.

Funding: British Council Assessment Research Grant 2017-2018, Project Title: "Applying Automated Analyses Techniques to Investigate Discourse Features in the APTIS for TEENS Writing Test: Evidence from Lower Secondary EFL Students", British Council, UK.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the anonymity of processed data.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Text Variables List

Basic Text Information

V1. No. of words in text (tokens)
V2. Syllables per word
V3. Words per sentence
V4. Average number of characters per word
V5. Average number of syllables per word
V6. No. of sentences (T-Units)
V7. Sentences per paragraph
V8. Average number of Sentences
V9. Text Easability-Narrativity
V10. Text Easability-Syntactic Simplicity
V11. Text Easability-Referential Cohesion
V12. Text Easability-Deep Cohesion
V13. Text Easability-Verb Cohesion
V14. Text Easability-Connectivity
V15. Text Easability-Temporality
V16. Text Easability-Word Concreteness

Word Frequency Indices

V17. Log freq. content words
V18. Log min. freq. content words

Readability Indices

V19. Flesch Reading Ease Index
V20. Flesch-Kincaid Grade Level
V21. Coh-Metrix L2 Readability

Lexical Diversity Indices

V22. VOCD (Lexical Diversity)
V23. MTLD (Lexical Diversity)
V24. Lexical Density (content words/total)
V25. Type-Token Ratio (all words
V26. Verb Density
V27. Adverb Density

Text Abstractness Indices

V28. Concreteness content words
V29. Imagability for content words
V30. Familiarity for content words
V31. Age of acquisition for content words
V32. Noun hypernymy
V33. Verb hypernymy
V34. Noun & Verb hypernymy
V35. Content Words Polysemy

Syntactic Complexity Indices

V36. Higher level constituents
V37. Noun Phrase incidence
V38. Modifiers per Noun Phrase
V39. Words before main verb
V40. Negations
V41. Passive sentences
V42. Syntactic structure similarity (adjacent sentences)
V43. Syntactic structure similarity (across paragraphs)
V44. Syntactic structure similarity (within paragraphs)

V45. Conditional operators
V46. All connectives
V47. Additive connectives
V48. Temporal connectives
V49. Causal connectives
V50. Logical connectives
V51. Contrastive connectives
V52. Positive connectives
V53. Negative connectives

Cohesion & Coherence Indices
V54. Causal cohesion
V55. Causal content
V56. Intentional content
V57. Temporal cohesion
V58. Spatial cohesion
V59. Logical operators

Referential Cohesion Indices
V60. Anaphoric reference
V61. Adjacent anaphoric reference
V62. Argument overlap
V63. Adjacent argument overlap
V64. Stem overlap
V65. Adjacent stem overlap
V66. Content word overlap
V67. Pronoun ratio
V68. Personal pronouns
V69. Impersonal pronouns
V70. LSA for adjacent sentences
V71. LSA for all sentences
V72. LSA for all paragraphs
V73. LSA given/new sentences

Additional Text Variables
V75. 1st person singular pronouns
V76. 1st person plural pronouns
V77. 2nd person pronouns
V78. 3rd person singular pronouns
V79. 3rd person plural pronouns
V80. Function Words
V81. Past Tenses
V82. Present Tenses
V83. Future Tenses
V84. Agentless Passive Voice
V85. Gerund density
V86. Infinitive density
V87. Noun incidence
V88. Verb incidence
V89: Adjective incidence
V90: Adverb incidence

Note

[1] Short words are defined as those words that include up to 4 letters whereas long words include between 5 and 12 letters (Harley 2014; Weekes 1997).

References

Abbott, Robert, Virginia Berninger, and Michel Fayol. 2010. Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology* 102: 281–98. [CrossRef]

Ai, H., and X. Lu. 2013. A corpus-based comparison of syntactic complexity in nns and ns university students' writing. In *Automatic Treatment and Analysis of Learner Corpus Data*. Edited by A. Diaz-Negrillo, N. Ballier and P. Thompson. Amsterdam: John Benjamins, pp. 249–64.

Banerjee, Jayanti, Florencia Franceschina, and Anne Margaret Smith. 2007. *Documenting Features of Written Language Production Typical of Different IELTS Band Score Levels*. IELTS Research Reports Volume 7. Canberra: IELTS Australia Pty and British Council.

Bardovi-Harlig, Kathleen, and Theodora Bofman. 1989. Attainment of Syntactic and Morphological Accuracy by Advanced Language Learners. *Studies in Second Language Acquisition* 11: 17–34. [CrossRef]

Beaman, Karen. 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In *Coherence in Spoken and Written Discourse*. Edited by Deborah Tannen. Asbury Park: ALEX Publishing, pp. 45–80.

Berninger, Virginia, Ana Cartwright, Cheryl Yates, Lee Swanson, and Robert Abbott. 1994. Developmental skills related to writing and reading acquisition in the intermediate grades. *Reading and Writing* 62: 161–96. [CrossRef]

Berninger, Virginia, Donald Mizokawa, and Russell Bragg. 1991. Theory-based diagnosis and remediation of writing disabilities. *Journal of School Psychology* 29: 57–79. [CrossRef]

Britton, Bruce, Glynn Shawn, Bonnie Meyer, and Muj Penland. 1982. Effects of text structure on the use of cognitive capacity during reading. *Journal of Educational Psychology* 74: 51–61. [CrossRef]

Brown, Cati, Tony Snodgrass, Susan Kemper, Ruth Herman, and Michael Covington. 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods* 40: 540–45. [CrossRef]

Bulté, Bram, and Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26: 42–65. [CrossRef]

Cacciari, Cristina, and Sam Glucksberg. 1995. Understanding idioms: Do visual images reflect figurative meanings? *European Journal of Cognitive Psychology* 7: 283–305. [CrossRef]

Caron, Jean, Hans Christoph Micko, and Manfred Thüring. 1988. Conjunctions and the recall of composite sentences. *Journal of Memory & Language* 27: 309–23.

Carrell, Patricia. 1987. Readability in ESL. *Reading in a Foreign Language* 4: 21–40.

Charrow, Veda. 1988. Readability vs. comprehensibility: A case study in improving a real document. In *Linguistic Complexity and Text Comprehension*. Edited by Alice Davison and Georgia Green. Hillsdale: Lawrence Erlbaum, pp. 85–114.

Cobb, Tom. 2007. Computing the vocabulary demands of L2 reading. *Language Learning & Technology* 11: 38–63.

Cobb, Tom. 2010. Learning about language and learners from computer programs. *Reading in a Foreign Language* 22: 181–200.

Covington, Michael. 2007. *CPIDR 3.0 User Manual*. CASPR Research Report 2007-03. Athens: Artificial Intelligence Center, University of Georgia.

Crossley, Scott, and Danielle McNamara. 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing* 182: 119–35. [CrossRef]

Crossley, Scott, and Danielle McNamara. 2011. Text coherence and judgments of essay quality: Models of quality and coherence. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Edited by Laura Carlson, Christoph Hoelscher and Thomas Shipley. Austin: Cognitive Science Society, pp. 1236–41.

Crossley, Scott, and Danielle McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35: 115–35. [CrossRef]

Crossley, Scott, and Danielle McNamara. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* 26: 66–79. [CrossRef]

Crossley, Scott, Jerry Greenfield, and Danielle McNamara. 2008. Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly* 42: 475–92. [CrossRef]

Crossley, Scott, Kristopher Kyle, and Danielle McNamara. 2016. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing* 32: 1–16. [CrossRef]

Crossley, Scott, Max Louwerse, Philip McCarthy, and Danielle McNamara. 2007. A Linguistic Analysis of Simplified and Authentic Texts. *The Modern Language Journal* 91: 15–30. [CrossRef]

Crossley, Scott, Tom Salsbury, and Danielle McNamara. 2009. Measuring L2 Lexical Growth Using Hypernymic Relationships. *Language Learning* 592: 307–34. [CrossRef]

Crossley, Scott, Tom Salsbury, and Danielle McNamara. 2010. The development of polysemy and frequency use in English second language speakers. *Language Learning* 60: 573–605. [CrossRef]

Crossley, Scott, Tom Salsbury, Danielle McNamara, and Scott Jarvis. 2011. Predicting lexical proficiency in language learner texts using computational indices. *Language Testing* 284: 561–80. [CrossRef]

Cumming, Alister, Robert Kantor, Kyoko Baba, Keanre Eouanzoui, Usman Erdosy, and Mark James. 2005. *Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for the New TOEFL® ETS*. Research Report Series No. MS-30. Princeton: Educational Testing Service.

Douglas, Dan. 1981. An exploratory study of bilingual reading proficiency. In *Learning to Read in Different Languages*. Edited by Sarah Hudelson. Washington, DC: Center for Applied Linguistics, pp. 93–102.

Dufty, David, Arthur Graesser, Max Louwerse, and Danielle McNamara. 2006. Assigning Grade Levels to Textbooks: Is it just Readability? In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*. Edited by Ron Son. Mahwah: Lawrence Erlbaum Associates, pp. 1251–56.
Ellis, Nick, and Alan Beaton. 1993. Psycholinguistic determinants of foreign language vocabulary acquisition. *Language Learning* 43: 559–617. [CrossRef]
Espada-Gustilo, Leah. 2011. Linguistic Features that Impact Essay Scores: A Corpus Linguistic Analysis of ESL writing in three proficiency levels. *3L: The Southeast Asian Journal of English Language Studies* 171: 55–64.
Ferris, Dana. 1994. Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly* 282: 414–20. [CrossRef]
Field, John. 2004. *Psycholinguistics: The Key Concepts*. New York: Routledge.
Gee, Nancy, Douglas Nelson, and Daniel Krawczyk. 1999. Is the concreteness effect a result of underlying network interconnectivity? *Journal of Memory and Language* 40: 479–97. [CrossRef]
Geva, Esther. 1992. The role of conjunctions in L2 text comprehension. *TESOL Quarterly* 26: 731–45. [CrossRef]
Goodwin, Sarah. 2016. A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing* 30: 21–31. [CrossRef]
Gorin, Joanna. 2005. Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement* 424: 351–73. [CrossRef]
Grabe, William, and Robert Kaplan. 1996. *Theory and Practice of Writing: An Applied Linguistic Perspective*. New York: Longman.
Graesser, Arthur, Danielle McNamara, and Jonna Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher* 405: 223–34. [CrossRef]
Graesser, Arthur, Danielle McNamara, Max Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers* 36: 193–202.
Grant, Leslie, and April Ginther. 2000. Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing* 92: 123–45. [CrossRef]
Guo, Liang, Scott Crossley, and Danielle McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing* 18: 218–38. [CrossRef]
Haberlandt, Karl. 1982. Reader expectations in text comprehension. *Advances in Psychology* 9: 239–49.
Haberlandt, Karl, and Arthur Graesser. 1985. Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology* 114: 357–74. [CrossRef]
Halliday, Michael, and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.
Harley, Trevor. 2014. *The Psychology of Language: From Data to Theory*, 4th ed. New York: Psychology Press.
Harrison, Colin. 1999. Readability. In *Concise Encyclopedia of Educational Linguistics*. Edited by Bernard Spolsky. Oxford: Elsevier, pp. 428–31.
Haswell, Richard. 2000. Documenting Improvement in College Writing: A Longitudinal Approach. *Written Communication* 173: 307–52. [CrossRef]
Horning, Alice. 1987. Propositional Analysis and the Teaching of Reading with Writing. *Journal of Advanced Composition* 6: 49–64.
Kaup, Barbara. 2001. Negation and its impact on the accessibility of text information. *Memory & Cognition* 297: 960–67.
Kemper, Susan. 1987. Life-span changes in syntactic complexity. *Journal of Gerontology* 423: 323–28. [CrossRef]
Kintsch, Walter. 1988. The Role of Knowledge in Discourse Comprehension: A Construction Integration Model. *Psychological Review* 95: 163–82. [CrossRef]
Kirschner, Michal, Carol Wexler, and Elana Spector-Cohen. 1992. Avoiding Obstacles to Student Comprehension of Test Questions. *TESOL Quarterly* 26: 537–56. [CrossRef]
Knoch, Ute, Susy Macqueen, and Sally O'Hagain. 2014. *An Investigation of the Effect of Task Type on the Discourse Produced by Students at Various Score Levels in the TOEFL iBT® Writing Test ETS*. Research Report Series No. RR-14–43. Princeton: Educational Testing Service.
Kobayashi, Hiroe, and Carol Rinnert. 1992. Effects of first language on second language writing: Translation versus direct composition. *Language Learning* 42: 183–215. [CrossRef]
Kuiken, Folkert, and Ineke Vedder. 2017. Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing* 34: 321–36. [CrossRef]
Laufer, Batia, and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16: 307–22. [CrossRef]
Liu, Hugo. 2004. MontyLingua: An End-to-End Natural Language Processor with Common Sense [Computer Software and Documentation]. Available online: http://web.media.mit.edu/~{}hugo/montylingua (accessed on 23 March 2012).
Lorenzo, Francisco, and Leticia Rodríguez. 2014. Onset and expansion of L2 cognitive academic language proficiency in bilingual settings: CALP in CLIL. *System* 47: 64–72. [CrossRef]
Louwerse, Max. 2002. An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics* 12: 291–315. [CrossRef]
Louwerse, Max, Philip McCarthy, Danielle McNamara, and Arthur Graesser. 2004. Variation in language and cohesion across written and spoken registers. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Chicago: Lawrence Erlbaum Associates, pp. 843–48.

Lu, Xiaofei. 2011. A corpus–based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 4451: 36–62. [CrossRef]
Lu, Xiaofei, and Haiyang Ai. 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing* 29: 16–27. [CrossRef]
MacWhinney, Brian. 2000. *The Childes Project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum Associates.
MacWhinney, Brian, and Catherine Snow. 1990. The Child Language Data Exchange System: An update. *Journal of Child Language 17* 2: 457–72. [CrossRef]
Malvern, David, and Brian Richards. 1997. A new measure of lexical diversity. In *Evolving Models of Language: Papers from the Annual Meeting of the British Association for Applied Linguistics Held at the University of Wales, Swansea, September 1996*. Edited by Ann Ryan and Alison Wray. Clevedon: Multilingual Matters, pp. 58–71.
Malvern, David, and Brian Richards. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19: 85–104. [CrossRef]
Malvern, David, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Houndmills: Palgrave Macmillan.
Markopoulos, George, George Mikros, and George Broussalis. 2011. Stylometric profiling of the Greek legal corpus. Paper presented at the 10th International Conference of Greek Linguistics, Komotini, Greece, September 1–4.
Martinez, Ana. 2018. Analysis of syntactic complexity in secondary education ELF writers at different proficiency levels. *Assessing Writing* 35: 1–11. [CrossRef]
Mazgutova, Diana, and Judit Kormos. 2015. Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing* 29: 3–15. [CrossRef]
McCarthy, Philip, and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 422: 381–92. [CrossRef] [PubMed]
McCutchen, Deborah. 1986. Domain Knowledge and Linguistic Knowledge in the Development of Writing Ability. *Journal of Memory and Language* 25: 431–44. [CrossRef]
McKee, Gerard, David Malvern, and Brian Richards. 2000. Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing* 15: 323–37. [CrossRef]
McNamara, Danielle, and Walter Kintsch. 1996. Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes* 22: 247–88. [CrossRef]
McNamara, Danielle, Max Louwerse, Philip McCarthy, and Arthur Graesser. 2010a. Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes* 474: 292–330. [CrossRef]
McNamara, Danielle, Scott Crossley, and Philip McCarthy. 2010b. Linguistic features of writing quality. *Written Communication* 271: 57–86. [CrossRef]
McNamara, Danielle, Zhiqiang Cai, and Max Louwerse. 2011. Optimizing LSA measures of cohesion. In *Handbook of Latent Semantic Analysis*. Edited by Thomas Landauer, Danielle McNamara, Simon Dennis and Walter Kintsch. New York: Routledge, pp. 379–400.
Meara, Paul. 2005. Lexical Frequency Profiles: A Monte Carlo Analysis. *Applied Linguistics* 26: 32–47. [CrossRef]
Mikros, George. 2007. Stylometric experiments in Modern Greek: Investigating authorship in homogeneous newswire texts. In *Exact Methods in the Study of Language and Text*. Edited by R. Köhler, G. Altmann and P. Grzybek. Berlin: Mouton de Gruyter, pp. 461–70.
Mikros, George. 2009. *Content Words in Authorship Attribution: An Evaluation of Stylometric Features in a Literary Corpus. Issues in Quantitative Linguistics*. Edited by R. Köhler. Lüdenscheid: RAM-Verlag, pp. 61–75.
Mikros, George. 2011. Automatic authorship attribution in Greek blogs. Paper presented at the 10th International Conference of Greek Linguistics, Komotini, Greece, September 1–4.
Mikros, George, and Konstantinos Perifanos. 2011. Authorship identification in large email collections: Experiments using features that belong to different linguistic levels. Paper presented at the PAN 2011 Lab, Uncovering Plagiarism, Authorship, and Social Software Misuse Held in Conjunction with the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation, Amsterdam, The Netherlands, September 19–22.
Muncie, James. 2002. Process writing and vocabulary development: Comparing lexical frequency profiles across drafts. *System* 30: 225–35. [CrossRef]
Nagabhand, Saranya, P. Nation, and Margaret Franken. 1993. Can Text be too Friendly? *Reading in a Foreign Language* 92: 895–907.
Nation, Paul. 2001. Using small corpora to investigate learner needs: Two vocabulary research tools. In *Small Corpus Studies and ELT*. Edited by Mohsen Ghadessy, Alex Henry and Robert Roseberry. Amsterdam: John Benjamins, pp. 31–45.
Nation, Paul. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63: 59–82. [CrossRef]
Nation, Paul, and Karen Wang. 1999. Graded readers and vocabulary. *Reading in a Foreign Language* 12: 355–80.
Ortega, Lourdes. 2003. Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics* 24: 492–518. [CrossRef]
Perfetti, Charles. 1992. The representation problem in reading acquisition. In *Reading Acquisition*. Edited by Philip Gough, Linnea Ehri and Rebecca Treiman. Hillsdale: Lawrence Erlbaum Associates, pp. 145–74.
Rashotte, Carol, and Joseph Torgesen. 1985. Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly* 20: 180–88. [CrossRef]

Rayner, Keith, Alexander Pollatsek, Jane Ashby, and Charles Clifton. 2011. *The Psychology of Reading*, 2nd ed. New York: Taylor & Francis Ltd.
Rayner, Keith, and Barbara Juhasz. 2004. Eye movements in reading: Old questions and new directions. *European Journal of Cognitive Psychology* 16: 340–52. [CrossRef]
Richards, Brian, and David Malvern. 2007. Validity and threats to the validity of vocabulary measurement. In *Modelling and Assessing Vocabulary Knowledge*. Edited by Helmut Daller, James Milton and Jeanine Treffers-Daller. Cambridge: Cambridge University Press, pp. 79–92.
Salsbury, Tom, Scott Crossley, and Danielle McNamara. 2011. Psycholinguistic word information in second language oral discourse. *Second Language Research* 27: 343–60. [CrossRef]
Scardamalia, Marlene, and Pamela Paris. 1985. The function of explicit discourse knowledge in the development of text representations and composing strategies. *Cognition and Instruction* 2: 1–39. [CrossRef]
Schmitt, Norbert, and Paul Meara. 1997. Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition* 19: 17–36. [CrossRef]
Schwanenflugel, Paula, Steven Stahl, and Elisabeth McFalls. 1997. Partial Word Knowledge and Vocabulary Growth during Reading Comprehension. *Journal of Literacy Research* 29: 531–53. [CrossRef]
Silver, Clayton, Glenn Phelps, and William Dunlap. 1989. Baddeley's grammatical reasoning test: Active versus passive processing differences re- examined. *Language Testing* 6: 164–71. [CrossRef]
Webb, Stuart. 2010. Pre-learning low-frequency vocabulary in second language television programmes. *Language Teaching Research* 14: 501–15. [CrossRef]
Weekes, Brendan. 1997. Differential effects of number of letters on word and nonword naming latency. *Quarterly Journal of Experimental Psychology* 50: 439–56. [CrossRef]
Weigle, Sara. 2002. *Assessing Writing*. Cambridge: Cambridge University Press.
White, Sheida. 2011. *Understanding Adult Functional Literacy: Connecting Text Features, Task Demands and Respondent Skills*. New York: Routledge.
Wind, Stefanie A., Catanya Stager, and Yogendra J. Patil. 2017. Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with L1 and L2 writing assessments. *Assessing Writing* 34: 1–15. [CrossRef]
Witte, Stephen, and Lester Faigley. 1981. Coherence, Cohesion, and Writing Quality. *College Composition and Communication* 322: 189–204. [CrossRef]
Yang, Weiwei, Xiaofei Lu, and Sara Cushing Weigle. 2015. Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing* 28: 53–67. [CrossRef]
Yu, Guoxing. 2009. Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics* 31: 236–59. [CrossRef]
Zadeh, Abdollah. 2006. The Role of Textual Signals in L2 Text Comprehension. *ESP Malaysia* 12: 1–18.
Zareva, Alla. 2007. Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research* 23: 123–53. [CrossRef]