Research Article

# Linguistic Features of Writing Quality and Development: A Longitudinal Approach

Scott A. Crossley, *Georgia State University*
Minkyung Kim, *Nagoya University of Commerce and Business*

## Structured Abstract

- **Background:** In terms of understanding writing through a linguistic lens, a number of recent studies have focused on assessing writing quality, likely because of rising interest in automated essay scoring, the availability of large corpora of scored essays, and the development of advanced natural language processing tools that can assess linguistic features in texts quickly and accurately. Fewer studies, however, have focused on investigating writing development in terms of linguistic features. Many of the studies that do investigate writing development use cross-sectional data to better understand how writing changes as a function of age or grade level.
- **Literature Review:** Cross-sectional studies of writing development using linguistic features generally find that as writers advance, they begin to produce more advanced linguistic features at the lexical, syntactic, and discourse levels (Berninger et al., 1996; Berninger et al., 2011; Crossley, Weston, et al., 2011; Haswell, 1986; King & Rentel, 1979; Verhoeven et al., 2002; Wagner et al., 2011).
- **Research Questions:** The purpose of the current study is to examine both writing quality and writing development from a linguistic perspective. Thus, this study is guided by the following two research questions:

1. To what extent can expert ratings of student writing quality be predicted by different time points (pre and posttest), prompts, class levels, and linguistic features?
2. Are there differences over time in student production of linguistic features that predict expert ratings of student writing quality?

- **Methodology:** This study examines both writing quality and writing development from a linguistic perspective by analyzing a corpus of persuasive essays that were collected from college-level writers during a 16-week longitudinal study. Essays were scored and a predictive model of writing quality was developed using linguistic features, time (pretest and posttest), class level, and prompt. A second analysis examined how linguistic features changed between the pretest and the posttest.
- **Results:** The results indicated that 11 linguistic features were predictive of writing quality. Time, class level, and prompt were not predictive of writing quality when co-varied with linguistic features. Of the 11 features that were predictive of essay quality, five of them demonstrated difference between pre and posttest scores.
- **Discussion:** Understanding how linguistic features are predictive of writing quality and how they change over time as writers advance can provide us with important indications of writing success and better explain how linguistic features interact with this success. We found that essay quality was most strongly predicted by linguistic features related to content, lexical sophistication, global cohesion, syntactic complexity, and grammatical accuracy. When co-varied with linguistic features, time, class level, and prompt were not significant predictors of essay quality, although post-hoc analyses did find that essay scores differed among pretests and posttests, class levels, and prompts.
- **Conclusions:** This data builds on previous studies that have shown similar trends in predicting writing quality but adds additional levels of detail and a greater focus on longitudinal growth, where research is rare.

*Keywords:* natural language processing, writing analytics, writing development, writing quality

## 1.0 Background

There are a variety of approaches that can be used to better understand writing from a linguistic perspective, including assessing links between text features and writing quality and examining writing development across grade levels and time (i.e., cross-sectional and longitudinal studies).

Assessing links between text features and writing quality can inform our understanding of how language components in texts may influence expert ratings and can provide indications of linguistic differences between low- and high-quality writing samples (Crossley et al., 2015; McNamara et al., 2010). Such knowledge can be used to inform pedagogy, standardized assessment, and automated essay scoring systems. Cross-sectional and longitudinal writing studies can also provide important information about developmental trajectories during the process of learning how to write (Applebee, 2002), especially from a linguistics perspective (Kress, 1994; Myhill, 2009), allowing teachers and administrators to set milestones and develop informed expectations.

In terms of understanding writing through a linguistic lens, a number of recent studies have focused on assessing writing quality, likely because of rising interest in automated essay scoring (AES; Strobl et al., 2019), the availability of large corpora of scored essays (Blanchard et al., 2013; Ishikawa, 2013), and the development of advanced natural language processing (NLP) tools that can assess linguistic features in texts quickly and accurately (Crossley et al., 2016; Kyle & Crossley, 2015). Fewer studies, however, have focused on investigating writing development in terms of linguistic features. Many of the studies that do investigate writing development use cross-sectional data to better understand how writing changes as a function of age or grade level. These studies generally find that as writers advance, they begin to produce more advanced linguistic features at the lexical, syntactic, and discourse levels (Berninger et al.,1996; Berninger et al., 2011; Crossley, Weston, et al., 2011; Haswell, 1986; King & Rentel, 1979; Verhoeven et al., 2002; Wagner et al., 2011). However, cross-sectional studies have a number of limitations; chief among them is that they do not follow the same writers over time and instead sample from different writers at different levels, making it difficult to control for differences among writers. Longitudinal studies correct for this limitation by sampling writing from the same writer(s) over time. However, longitudinal studies require a greater number of resources and suffer from more participant attrition, potentially explaining why longitudinal studies of writing development are rare compared to cross-sectional studies. In particular, there have been few longitudinal studies that focus on linguistic development (Myhill, 2009) even with advances in corpus development (Myhill, 2008), mainly because large longitudinal writing corpora are not readily available.

The purpose of the current study is to examine both writing quality and writing development from a linguistic perspective. We do so by analyzing a corpus of persuasive essays that were collected from college-level writers during a 16-week longitudinal study in which pre and posttest essays were collected. These essays were then scored by expert raters, and a predictive model of writing quality was developed using linguistic features from the text (content, lexical sophistication, syntactic complexity, cohesion, and grammar accuracy), time (pretest and posttest), class level (e.g., Composition I and Composition II classes), and prompt. We followed this analysis up with analyses of how linguistic features predictive of essay quality changed between the pretest and the posttest.

College-level longitudinal writing studies that focus on linguistic development, such as this one, are uncommon, with only a few example studies known, all of which focus on syntactic and/or lexical changes over time. These studies generally indicate that, over time, college-level writers begin to produce more sophisticated words (Haswell, 2000; MacArthur et al., 2019) and potentially more complex syntactic features (Haswell, 2000). Longitudinal studies that include a writing quality component are even more rare, with only one known study (MacArthur et al., 2019).

## 2.0 Literature Review

### 2.1 Linguistic Features and Writing Analyses

There are many ways to assess writing development and quality, including disciplinary expertise, emergent patterns of writing skills, writers' responses and choices, writers' perceptions, social and psychological growth, and analysis of the language produced by writers (Gere, 2019). Studies that focus on how linguistic features found in student writing can inform investigations of text quality and writer development have been common since the 1970s. These studies demonstrate that linguistic patterns in texts are predictive of text quality and writing development. The most common linguistic constructs used in these analyses are related to lexical, syntactic, and cohesion features in the text (Berninger et al., 1992), with studies focusing on how these features work in isolation or in tandem with one another to predict quality or changes over time. The very basic notion underlying these studies is that more proficient writers will produce more sophisticated linguistic features. Additionally, as writers develop, they will begin to produce more sophisticated linguistic features (Crossley et al., 2011)

The most commonly reported features are strongly related to lexical features with sub-constructs related to lexical diversity (i.e., the number of unique words) and lexical sophistication (i.e., the intrinsic difficulty of words). Theoretically, the production of a greater number of different words or words that are more sophisticated is related to greater lexical acquisition. For instance, usage-based approaches to explaining language acquisition (Goldberg, 2006) indicate that the production of more sophisticated words is related to greater exposure to words, associative learning, automatization, and representations of word meaning and form (Langacker, 2007). Psycholinguistic studies have also shown that more sophisticated words are more difficult to recognize and process (Balota et al., 2007).

In terms of measuring lexical features in texts, lexical diversity is traditionally associated with type-token (TTR) ratios that calculate the number of unique words in a text divided by the total number of words in a text (Bates et al., 1988). Lexical sophistication has traditionally been operationalized through frequency metrics (i.e., how common a word is in a language), with more infrequent words being more sophisticated or difficult (e.g., Brysbaert et al., 2018; Laufer & Nation, 1995). However, with time, the definition of lexical sophistication has expanded to include academic words (Coxhead, 2000), words that are spelled (orthographic) and pronounced (phonological) like other words (i.e., word neighborhood effects), words that take longer to name (Balota et al., 2007), words that are more specific (Fellbaum, 1998), and words that have more difficult lexical properties in terms of concreteness, imageability, and familiarity (Stadthagen-Gonzalez & Davis, 2006).

A second common linguistic feature that is analyzed in writing studies is syntactic complexity, which refers to the complexity and variety of syntactic forms used by a writer (Beers & Nagy, 2009). Just as the production of more advanced lexical items is related to greater lexical acquisition, the production of more advanced syntactic features indicates a greater knowledge of a language's syntactic structure (e.g., Jagaiah et al., 2020). Generally, movements from less complex to more complex syntax can be explained by syntactic theories and hypotheses. For example, the implicit learning theory assumes that the acquisition of complex syntactic structures takes place via an error-based implicit learning mechanism and through meaning-form mappings (e.g., Chang et al., 2006). Of particular interest for the context of college-level writing is the sequencing of syntactic development in academic writing proposed by Biber et al. (2011). Biber and colleagues propose that novice academic writers primarily produce greater clausal complexity features (e.g., finite dependent clauses) common in speech, and, as they develop, they gradually make greater use of phrasal complexity features (e.g., dependent phrases and noun phrase modifiers) that are more strongly related to advanced academic writing (Staples et al., 2016).

Historically, syntactic analyses of text production focused on sentence length and T-units (i.e., dominant clauses and all subordinate clauses) to assess syntactic complexity. More novel approaches include calculating phrasal and clausal complexity (i.e., the number of modifiers in a noun phrase or the number of words before the main verb of a sentence; Crossley & McNamara, 2014) or assessing the frequency of verb-argument constructions (Kyle & Crossley, 2017). There have also been recent moves to examine more phrasal components of writing because phrases comprise both lexical and syntactic features (Sinclair, 1991) and phrasal knowledge is an important component of linguistic ability (Ellis, 2012; Siyanova-Chantura & Martinez, 2015).

Beyond lexical and syntactic features, writing researchers have also considered discourse-level features, usually in terms of text cohesion. Text cohesion moves beyond the word and sentence level and examines the inter-connectivity of text segments in writing, which can be indicative of lexical, semantic, and argumentative dependencies within a text (Halliday & Hasan, 1976). Text cohesion can occur at the sentence level (i.e., local cohesion), among larger segment gaps such as paragraphs and chapters (i.e., global cohesion), and among texts (e.g., inter-document cohesion). Historically, text cohesion was analyzed at the local level through explicit links between text segments either through the use of pronouns to reference previously mentioned elements, the repetition of lexical items, and/or the use of connectives to link ideas together. Text cohesion is learned, and good writers can maintain cohesion in a text by allowing readers to better understand and evaluate relations in the text and develop a mental representation of that text (i.e., text coherence; McNamara et al., 1996; Sanders & Pander Maat, 2006).

To a lesser degree, researchers have also been interested in assessing links between writing quality and development and grammatical and mechanical errors (e.g., spelling and punctuation errors) because effective writing depends on knowledge of the language system to include grammatical and mechanical rules. The importance of these rules is evidenced in writing literature that highlights their importance in explaining writing quality (Eckes, 2008; Santos, 1988; Zhu, 2004), the importance of teaching the rules to students (Graham, 1983; Morris et al., 1995), and teacher beliefs about the importance of these rules (Cutler & Graham, 2008).

## 2.2 Writing Quality

A number of studies have examined links between linguistic features in text and human ratings of essay quality in order to better understand how linguistic features equate to writing proficiency. Lexically, studies have demonstrated that more proficient first-year undergraduate writers use more academic words in expository writing (Douglas, 2013) and produce more sophisticated phrasal items, including less frequent phrases in argumentative writing (Crossley et al., 2012). Syntactically, studies have shown that school-aged students that produce more complex syntactic structures are judged to be more proficient writers (Benson & Campbell, 2009; Jagaiah et al., 2020; Klecan-Aker & Hendrick, 1985; Myhill, 2009). For example, Klecan-Aker and Hendrick (1985) reported that ninth graders tended to produce a greater number of words per T-units and clauses than sixth graders. Myhill (2008) found that for eighth and tenth graders, higher-rated narratives tended to include a greater number of finite verbs, coordinated clauses, and subordinated clauses. However, not all studies report strong associations between essay quality and syntactic complexity. As an example, McNamara et al. (2013) found that for first-year argumentative essays, syntactic features including incidences of verb phrases and modifiers per noun phrase were significantly correlated with essay quality, but these features failed to add a significant contribution to essay quality over and above lexical sophistication and cohesive features. Similarly, Perin and Lauterbach (2018) found that for community college students, higher-rated persuasive essays did not differ in syntactic complexity (e.g., number of words before the main verb) from lower-rated ones.

In terms of cohesion, school-aged writers that produce more markers of cohesion (Cameron et al., 1995; Struthers et al., 2013) and use them appropriately (Cox et al., 1990) are reported to be better writers. However, differences may exist in the production of connectives, with higher-quality samples containing more additive, manner, causal, and adversative adverbs but fewer temporal adverbs (Myhill, 2008). Links between cohesion features and writing quality are mixed for college-level writers, especially for local cohesion devices. For instance, Witte and Faigley (1981) and MacArthur et al. (2019) reported that a greater density of cohesive ties and referential cohesion (i.e., lexical connections across sentences) was related to higher scores for college writers, respectively. However, Perin and Lauterbach (2018) reported negative relationships between referential cohesion and essay quality for community college students. Other research in college-level writing has reported no links or negative correlations between local cohesive features including many different types of connectives (e.g., logical operators, positive logical connectives, and negative temporal connectives) and sentence-level word and semantic overlap features (Crossley & McNamara, 2010, 2011; Evola et al., 1980; McCulley, 1985; McNamara et al., 2010; Neuner, 1987). This contrasts with measures of global cohesion, which report strong links with text quality (Neuner, 1987). For example, work by Crossley and colleagues has demonstrated that for college-level writing, markers of global cohesion that measure lexical and semantic similarity across paragraphs are positively correlated with human ratings of essay quality (Crossley & McNamara, 2010, 2011, 2016; Crossley, Roscoe, et al., 2011).

Empirical studies linking grammar errors and writing quality are rare, with at least one study (Crossley et al., 2014) reporting that grammatical errors in essays only had a small effect on expert judgments. Stronger links have been reported between mechanical errors and essay quality for school-aged students (Morris et al., 1995). For example, Graham et al. (1997) found that mechanics errors accounted for a significant portion of the variance for both text fluency and essay quality scores for narrative and expository writing in first through sixth grade. In a more recent study, Crossley et al. (2014) reported that spelling errors yielded the strongest correlations with human judgments of essay quality for high school students.

There have also been a number of studies that combine multiple linguistic features to develop models of writing quality or to classify low- and high-quality essays. In an early study, McNamara et al. (2010) examined differences between low- and high-quality undergraduate essays using three linguistic features: number of words before the main verb (syntactic complexity), lexical diversity (both cohesion and lexical sophistication), and word frequency. These three features accurately classified 67 percent of the essays as being either low or high quality and predicted 22 percent of the variance in essay scores. In a later study, using a larger corpus of 997 persuasive essays written by ninth, tenth, and twelfth graders and college freshmen, Crossley et al.(2015) predicted 52 percent of the variance in essay scores using a variety of linguistic features, including text length (i.e., essay and paragraph length) and lexical variables (nominalizations, lexical diversity, word frequency, and word specificity). In a recent study of college writers, MacArthur et al. (2019) examined relations between linguistic features and writing quality and differences in linguistic features in a pre- and a post-writing task for 252 college-level writers across a semester of study in both control and experimental conditions. They found that text length, cohesion, syntax, and word-level features were significant predictors of writing quality and explained ~50 percent of the variance in essay scores.

## 2.3 Writing Development

A number of previous studies have shown clear trends in writing quality as a function of time. For instance, Berge et al. (2019) examined changes in writing quality for 3,088 third through seventh grade students over two years in 24 schools across Norway and found that primary school students' writing quality improved significantly across that time. However, they did report differences in improvements across schools, classes, and students.

Studies have also shown trends in the development of linguistic features over time using two different approaches: cross-sectional and longitudinal. Cross-sectional approaches generally examine texts from a number of different writers all sampled once. Cross-sectionally, development can be traced by examining differences in writers by grade level (e.g., differences between sixth and ninth grade writers), age, or proficiency level. Cross-sectional analyses allow for generalization about development across larger samples of writers, but they do not track development at the individual level (and thus cannot attend to potential differences between writers). Longitudinal approaches, on the other hand, track individual development by collecting writing samples from the same writers at different time points and examining how writers change over time. Longitudinal approaches allow for greater control of individual differences between writers because the data is repeated.

Previous cross-sectional studies have found differences in linguistic production across grade levels for lexical, syntactic, and cohesive features. In an early study, Berman and Verhoeven (2002) found increases in mean length of clause as a result of grade level, with later grades producing longer clauses. They also reported increased lexical diversity (as measured by voc-D; Malvern & Richards, 2002) as a function of age, especially between junior high school and high school students, indicating that older students used a greater variety of words (lexical sophistication) and did not repeat as many words (lower text cohesion). They concluded that changes in text construction occur as students advance in grades and these changes go hand in hand with changes in psycholinguistic and socio-cultural functioning. Another study that examined multiple linguistic constructions (Crossley, Weston, et al., 2011) investigated differences between ninth grade, eleventh grade, and college-level writers. Crossley, Weston, et al. (2011) found that the strongest predictors of students' grade level were lexical features such as word frequency, word concreteness, and word polysemy. These features changed such that writers at higher grade levels produced more infrequent words, more concrete terms, and words with fewer senses (i.e., more sophisticated words). Crossley and colleagues also reported difference by grade level for cohesion features (incidence of positive logical connectives and type-token ratio, which measures word repetition) and a measure of syntactic complexity (number of modifiers per noun phrases) such that more advanced writers used fewer connectives, repeated words less often and had more complex syntactic structures.

A number of cross-sectional studies have focused on differences in syntactic complexity across grades. For instance, children demonstrate growth in complete sentence use at the expense of run-on sentences and sentence fragments, which decrease over time (Berninger et al., 2011). Much cross-sectional research focuses specifically on T-unit features. Such research reports increased T-unit complexity as a function of grade level (between first, fourth, eighth, and twelfth grades; Haswell, 1986; Hunt, 1965, 1966, 1970; Wagner et al., 2011). One problem with T-units is that they do not provide information about what syntactic elements in a sentence lead to longer or shorter T-units. In response, studies have examined the specific syntactic features that lead to longer T-units as a function of grade level, reporting that as writers advance, they produce T-units that contain a greater number of relative clauses, complement clauses, subordinate clauses, infinitives, passives, and modals as well as a wider variety of clause types and longer noun phrases (Berninger et al., 2011; Haswell, 1986, 1990; Perera, 1984; Verhoeven et al., 2002). Few studies have focused on the development of grammar and mechanical knowledge, with at least one study (Wharton-McDonald et al., 1998) reporting that essays written by higher proficiency students were more accurate in terms of mechanics.

Cross-sectional studies have also focused on the development of cohesion features over time, generally finding that writing moves from more local cohesion at lower grades to greater global cohesion at higher grades. As an example, younger students connect ideas at the sentence level (i.e., local cohesion; Berninger et al., 1996) through features such as pronoun repetition and the use of connectives (King & Rentel, 1979). Students continue to progress in their use of local cohesion devices until around the eighth grade when production seems to level off (McCutchen & Perfetti, 1982). While younger writers produce more local cohesion devices over time, older writers develop cohesion strategies that link ideas and topics across paragraphs (i.e., global cohesion; Bereiter & Scardamalia, 1987; Flower & Hayes, 1980). However, there is also evidence that as older students move away from explicit cohesion devices like the use of connectives, they begin to use more complex syntactic features to develop text cohesion (Haswell, 2000; McCutchen & Perfetti, 1982).

Fewer researchers have examined writing development longitudinally across multiple participants. A large-scale analysis conducted by Loban (1976) focusing on development from kindergarten to high school found that over time, writers showed syntactic changes such that they begin to produce longer sentences with a greater number of embedded clauses and longer noun phrases. Myhill (2009) reported similar findings in her study of secondary students in that older students wrote longer sentences that had a greater number of embedded phrases and more subordination in those sentences. For college-level writers, Haswell (2000) examined growth in 64 students between the first and third semester of college. Haswell found an increase in long words (greater than nine letters) over time and reported that students began to produce longer sentences with longer clauses. MacArthur et al. (2019) reported that lexical complexity scores increased in posttest essays but reported no differences for syntactic complexity features for college-level students.

## 3.0 Research Questions

The current study seeks to build on the recent study by MacArthur et al. (2019) by examining both writing quality and writing development at the college level in terms of linguistic features. We do so by investigating scored writing samples collected in a standardized assessment at the beginning and end of semester-long classes. Unlike MacArthur et al.'s (2019) study, there were no experimental conditions, so the study is longitudinal in nature without treatment designs (although we did sample from three different levels of composition classes: Composition I, Composition II, and Advanced Composition). Also, unlike MacArthur et al., our focus was on individual linguistic features (as compared to macro-features), so we examined micro-features of language and did not include structural components such as text length in our analyses. Our statistical analyses also differed in that we used linear mixed effects (LME) models to predict essay scores while including co-varying effects like time and class and controlling for random effects including participants and teachers. In this way, we can analyze the predictive strength of not only linguistic features but non-linguistic features in predicting essay quality. We follow our LME model with an analysis of difference over time for linguistic features that were predictive of essay quality to see if student development matches scoring expectations. Thus, this study is guided by the following two research questions:

1. To what extent can student writing scores in a standardized assessment be predicted by different time points (pre and posttest), prompts, class levels, and linguistic features?
2. Are there differences over time in student production of linguistic features that predict student writing scores in a standardized assessment?

## 4.0 Method

### 4.1 Corpus

Our corpus comprised 613 essays written by 313 undergraduate students who took first-year composition courses at a southeastern university in the United States. The university was a large state university with a population of mostly White students (~80%). The second largest group of students was African American (~18%), followed by Hispanic students (~2%). The original purpose of the data collection was to internally assess the effectiveness of the university's freshman composition program. Within this narrow mandate, no individual difference or demographic data was collected from students.

The students took one of three composition classes offered by the English Department: Composition I, Composition II, and Advanced Composition. Composition I classes are generally taken by first-year students in their first semester, while Composition II classes are taken by second-semester students. Both Composition I and II are required courses. Advanced Composition classes were available for students who performed well on the written portion of the university’s entrance exam. Students that successfully passed Advanced Composition did not need to complete Composition I and II. . Non-native speakers of English were required to take separate composition classes, helping to ensure that the entire population in this sample consisted of native speakers.

Throughout the semester, students were expected to regularly complete writing assignments, including five major assignments that were common across the three different composition classes. The topic of these assignments differed by instructor, but the general expectations were the same, with all students expected to produce an argumentative analysis paper based on assigned readings, a controlled research project in which they selected a topic and synthesized information into an argumentative essay, a short fiction analysis, and a poetry research piece. Students also had a final exam with timed in-class writing on an integrated writing prompt that included around four information sources. All students, regardless of class, were also introduced to MLA style expectations, writing strategies, library research, peer reviewing, and literary genres.

In addition, students produced two essays in two timed sessions of their composition classes: one at the beginning of the semester and the other at the end. In this study, we used a total of 500 essays written by 250 students who completed both sessions (all the students who completed both the pre and posttest writing samples). The students were given 25 minutes to write each essay, with no outside referencing allowed. Two SAT writing prompts were used in the corpus collection. One prompt was about originality and uniqueness, while the other was about admiring heroes versus celebrities. Both prompts, along with their writing assignments, can be found in Appendix A. The prompts were used in retired SAT tests taken by high school students attempting to enter post-secondary institutions in the United States. Thus, the writing samples represent a specific type of writing: writing found in standardized assessments. Prompts were counterbalanced, but due to sampling issues, around two-thirds of participants (*n* = 156) wrote an essay about originality in the first session and then an essay about admiration in the second session, while the other third (*n* = 94) did the opposite. Sampling issues included different numbers of students in classes (all students in a single class had the same prompt assigned to them at pretest and posttest) and an instructor who did not collect posttest essays for her two classes.

### 4.2 Essay Ratings

Essays were scored by trained raters on overall quality using a holistic, six-point grading scale, which was a standardized rubric commonly used in assessing SAT essays (see Appendix B). The scale focuses on test-takers' development of a point of view on the issue, critical thinking, use of appropriate examples, accurate and adept use of language, use of a variety of sentence structures, and errors in grammar and mechanics as well as text organization and coherence. Each essay was read by two raters. The raters had either a master's or a doctoral degree in English and at least two years of experience teaching composition classes at the university level. All raters were full-time faculty within the English Department. For training purposes, the raters first scored 20 practice essays that were not included in the corpus. Pearson correlations were calculated to measure inter-rater reliability. After an inter-rater reliability of Kappa of at least .60 was reached in the training set, the raters scored the essays in the corpus independently. Initial inter-rater reliability after score was Kappa = .665. If score differences between two raters were two points or greater, the raters adjudicated the final score through discussion. If agreement was not reached, the score was not changed. Average scores between the raters for the adjudicated scores were calculated for each essay and used for the data analysis.

### 4.3 Assessment of Linguistic Features

To examine the relationship between essay quality and linguistic variables, five types of linguistic features found in student writing as informed by theoretical and data-driven accounts of language acquisition, production, and writing development were considered: lexical sophistication and diversity, syntactic sophistication and complexity, grammar and mechanics, cohesion, and content. Not only was the selection of these features informed by previous theoretical and data-driven research, all of the selected linguistic features were also in line with the scoring rubric used in this study. Additionally, since our interest was in linguistic features, we did not consider text structures such as number of words, sentences, or paragraphs, which are more strongly associated with fluency and knowledge of text structure. To measure lexical sophistication and lexical diversity, the Tool for the Automatic Analysis of Lexical Sophistication (TAALES 2.2; Kyle & Crossley, 2015; Kyle et al., 2018) and the Tool for the Automatic Analysis of Lexical Diversity (TAALED 1.2.4) were used, respectively. Syntactic features were measured using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC 1.3.8; Kyle, 2016), while grammar and mechanics features were assessed using the Grammar and Mechanics Error Tool (GAMET 1.0; Crossley et al. , 2019). Cohesive features were assessed using the Tool for the Automatic Analysis of Cohesion (TAACO 2.0.4; Crossley et al., 2016), while, for content analysis, a measure of differential word use (DWU), which distinguishes word use of higher-rated essays from that of lower-rated essays based on essay content, was calculated following Attali (2011). Each tool or measure used for assessing linguistic features is briefly discussed in the following sections, along with the theoretical and data driven reasons for its selection.

**4.3.1 TAALES.** TAALES (Kyle & Crossley, 2015; Kyle et al., 2018) measures approximately 400 lexical and phrasal features. The features we selected were specifically related to lexical sophistication, which has important theoretical links to usage-based theories of language acquisition (Ellis, 2002) and strong overlap with psycholinguistic studies of language acquisition (Balota et al., 2007). Additionally, many of the selected features have shown predictive strength in previous writing development of quality studies (e.g., Crossley et al., 2012; Douglas, 2013). The features we selected included the following

- lexical frequency (i.e., scores based on how often a word occurs in a reference corpus)
- lexical range (i.e., scores based on how many documents in a reference corpus include a word)
- psycholinguistic word information (e.g., human ratings of familiarity)
- semantic relations (e.g., hypernymy and polysemy, which measure specificity and abstractness respectively)

- n-gram (i.e., sequences of contiguous words)
- frequency (i.e., scores based on how often an n-gram occurs in a reference corpus)
- n-gram range (i.e., scores based on how many documents in a reference corpus contain an n-gram)
- n-gram association strength (i.e., how strongly a combination of words is attached to each other)
- academic language (i.e., lexical and phrasal items that occur frequently in an academic corpus)
- age of exposure (i.e., the age at which individuals are likely to be first exposed to a certain word ).

In calculating indices related to frequency, range, and association strength, various reference corpora are used, such as the Thorndike-Lorge Corpus (Thorndike & Lorge, 1944), the SUBTLEXus corpus of subtitles (Brysbaert & New, 2009), or/and the Corpus of Contemporary American English (COCA; Davies, 2009). N-gram association strength is measured by various indices including Mutual Information (MI; how often sequences of contiguous words co-occur in a reference corpus) and Approximate Collexeme (AC; joint probability which two words will co-occur).

**4.3.2 TAALED.** To examine features related to lexical knowledge, we investigated indices related to lexical diversity (i.e., calculations of the number of unique words produced) using TAALED. TAALED calculates approximately 30 lexical diversity indices, including the measure of textual lexical diversity (MTLD; the mean length of sequential word strings in a text that maintain a given TTR value, such as .720; McCarthy & Jarvis, 2010), hypergeometric distribution diversity (HD-D; the probability of encountering any of its type in a random sub-sample drawn from the text; McCarthy & Jarvis, 2010), and Maas (i.e., log corrections of TTR). Indices are calculated using all lemmas, content lemmas, or function lemmas.

**4.3.3 TAACO.** We selected a number of cohesion indices from TAACO (Crossley et al., 2016) that overlap with measures of text cohesion that examine connectivity between segments of text (Halliday & Hasan, 1976) including features of local and global cohesion. TAACO computes around 150 indices of text cohesion. We selected indices related to repetition of words throughout a text calculated using type-token ratios (TTRs; the number of unique lemmas [types] divided by the number of total running lemmas [tokens]) for all words, part of speech tags, content words, and function words. We also selected overlap of all words, part of speech tags (e.g., nouns and pronouns), content words, and function words across adjacent sentences and adjacent paragraphs. Lastly, we selected connective indices including logical connectives (i.e., words that connect sentences, such as moreover and nevertheless) and temporal connectives (i.e., words that connect sentences in time, such as after and before).

**4.3.4 TAASSC.** We also selected a number of syntactic features that tap into the complexity of structures used by writers as proxies for writers' knowledge of syntactic structures. We did so using TAASSC (Kyle, 2016), which measures approximately 370 indices related to clausal and phrasal indices of syntactic complexity and indices related to complexity of verb-argument constructions (defined as a main verb plus all of its direct dependents)[1]. We specifically selected measures of clausal complexity, such as average numbers of particular structures per clause and dependents (e.g., passive agents) per clause, and noun-phrase complexity features based on the average number and standard deviations of dependents (e.g., determiners relative clause modifiers) per noun. We also selected indices from TAASSC that calculated verb-argument construction complexity, such as the frequency of verb-argument constructions, TTR of verb-argument constructions, and the strength of association between the verb-argument construction and the main-verb lemma. Lastly, we included indices from Lu's (2010) Syntactic Complexity Analyzer, such as mean length of sentence and clauses per sentence.

**4.3.5 GAMET.** To measure writers' grammatical and mechanical knowledge, we included features derived from GAMET (Crossley et al., 2018), which assesses approximately 290 indices related to grammatical and mechanics errors in texts. We selected five types of errors: grammatical errors (e.g., subject-verb agreement, sentence fragments, plurals, and verb tense), spelling errors, punctuation errors (e.g., two consecutive periods and the use of apostrophes), white space errors (i.e., lack of a space between sentences), and repetition errors.

**4.3.6 DWU Measure.** We used differential word use (DWU) to measure content of the writers' texts. DWU compares each word's relative frequency in high-quality essays with that in low-quality essays. Words with higher DWU scores indicate those which appear more frequently in high-scored essays. The assumption is that words that are more frequently used in higher-rated essays than in lower-rated essays are indicative of vocabulary more typical of high proficiency writing across different prompts in the given writing task (Attali, 2011). While the DWU measure can be interpreted as a lexical measure, it is a more likely content measure because it is based on actual essay content produced by test-takers.

### 4.4 Statistical Analysis

We first conducted correlation analyses to identify linguistic features which showed significant correlations with holistic essay scores. Linguistic variables which showed correlations with

[1] For example, in the sentence *Mary likes dogs*, the main verb *likes* takes two direct dependents: *Mary* and *dogs*. This clause is represented by the verb-argument construction (VAC) *nominal subject – verb – direct object*.

scores higher than |.200| were retained. For those variables, we controlled for multicollinearity to ensure we did not include indices that measured similar language features. We did so using variance inflation factors (VIF) such that any variables showing a VIF above five were deleted (Fox & Weisberg, 2010). This left us with 47 linguistic variables.

To predict holistic writing scores (Research Question 1), we used a linear mixed effects (LME) modeling approach. LME models take into account fixed effects (i.e., variables predicting our dependent variable: essay score) and random effects (i.e., variables that control for random variance in the data). We used LME models to control for repeated writing samples per individual, such that baseline writing scores (i.e., intercepts) for each individual were represented as a random effect in the LME model (Barr et al., 2013). In addition, teacher was also added as a random factor to control for the teacher effect. Thus, in the random intercept model, a different baseline value for writing scores was assigned to each participant and to each group of students who were taught by the same teachers. Given this random intercept model, linguistic features were added as fixed effects. For the linguistic variables, the LME model was developed by backward selection of the fixed effects using log-likelihood ratio tests to select the fixed effects that reached the significance level ($t > 1.960$ at a .050 significance level). Time (i.e., pretest and posttest essay collections), prompts, and class levels (i.e., Composition I, Composition II, and Advanced Composition) were also added as fixed effects. Time was added as a fixed effect because it was expected that participants' writing scores would increase over time. Prompts were added as a fixed effect because there may be a prompt effect on writing scores. Class levels were added as a fixed effect because students from different class levels may perform differently. Additionally, interaction effects between time and class levels, time and linguistic features, and time and prompt were added. The latter interaction was included to test whether issues in counter-balancing influenced outcomes. We then added a random slope adjustment for the teacher factor for each class because the class effect on writing scores may differ depending on the teacher.[2] Thus, the final LME model included the linguistic variables along with time, prompts, and class levels as fixed effects, and participants and teachers as random factors.

To answer Research Question 2, paired *t* tests were conducted. The *t* tests were used to examine differences over time in linguistic features. However, we only conducted paired *t* tests for the linguistic features that predicted writing scores in the LME model. We conducted the paired *t* tests to examine if writers showed changes in their linguistic features as a function of time spent in the classroom.

[2] While it is generally recommended to consider interaction and random slope effects in LME models in experimental settings (Barr et al., 2013), we did not consider either interactions among linguistic features or random slopes. Interaction effects among linguistic features were not of interest because it is expected that there are certain levels (however small) of interactions among the linguistic features included in the model as these features were measured using the same text. Thus, it would be almost impossible to consider all of the potential interaction effects among the linguistic features. Also, we did not consider a random slope model for linguistic features. This is because effects of computational linguistic measures (i.e., fixed effects) on human rating of writing scores (i.e., independent variable) are not likely to differ depending on the subject factor or the teacher factor (i.e., random factors).

For data analysis, we used R (R Core Team, 2016) and various R packages. The *lme4* package (Bates et al., 2015) was used to construct LME models. The *LMERConvenienceFunctions* package (Tremblay & Ransijn, 2015) was used to perform backward selection of fixed effects. The *lmerTest* package (Kuznetsova et al., 2016) was used to calculate *p* values from the models. Finally, the *MuMIn* package (Bartoń, 2017) was used to calculate two measures of variance explained from the models (i.e., a marginal *r*-squared that calculated the variance explained by the fixed effects only, and a conditional *r*-squared that calculated the variance explained by both the fixed and random effects).

## 5.0 Results

### 5.1 Descriptive Statistics

Table 1 shows the number of essays, the mean score, and the mean number of words across different time points, classes, and prompts.

**Table 1**

*Mean Scores and Word Counts Across Different Classes, Times, and Prompts*

| Class level | Time | *N* | Mean score (*SD*) | Mean word counts (*SD*) |
|---|---|---|---|---|
| Composition I | Time 1 | 96 | 3.047 (.883) | 305.260 (107.962) |
| Composition I | Time 2 | 96 | 3.120 (.888) | 341.438 (115.409) |
| Composition II | Time 1 | 82 | 3.140 (1.001) | 332.366 (113.630) |
| Composition II | Time 2 | 82 | 3.500 (1.097) | 445.342 (143.891) |
| Advanced Composition | Time 1 | 72 | 3.542 (1.003) | 368.500 (123.468) |
| Advanced Composition | Time 2 | 72 | 3.701 (.834) | 394.361 (118.568) |
| All classes | Time 1 | 250 | 3.220 (.976) | 332.364 (116.854) |
| All classes | Time 2 | 250 | 3.412 (.975) | 390.760 (133.222) |
| Prompt 1 (originality) | Times 1-2 | 250 | 3.222 (.973) | 346.132 (121.636) |
| Prompt 2 (admiration) | Times 1-2 | 250 | 3.410 (.978) | 376.992 (133.572) |
| All classes | Times 1-2 | 500 | 3.320 (.980) | 361.560 (128.550) |

### 5.2 LME Model Construction

As a baseline model, a random intercept model without fixed effects was constructed by including the participant and teacher factors as random intercepts. This model explained 38.955 percent of the variance in holistic writing scores. Based on the random intercept model, an LME model was created using backward selection, which included 11 linguistic features as significant fixed effects. We then added three additional fixed effects (i.e., time, prompts, and class levels) to the final LME model, but none of these three fixed effects were significant. No interaction effect between class levels and time or between linguistic features and time was revealed.

The final LME model is shown in Table 2. In the final model, the fixed effects explained 58.960 percent of the variance in writing scores, while both the fixed and random effects explained 61.600 percent of the variance in writing scores. When considered together with the fixed effects, the random effects of participants and teacher were small, explaining 3.065 percent of the variance in writing scores. The model included 11 significant linguistic variables including content-related, lexical, syntactic, cohesive, and grammatical error features. With respect to content features, results indicated that essays with higher scores tended to include words that appeared in higher scoring essays as assessed by the DWU measure. Five lexical features significantly predicted writing scores, such that higher-rated essays tended to include words that are learned at a later age, function words that occur less frequently in the COCA fiction corpus, content words that occur less frequently in the COCA academic corpus, bigrams with stronger associations as measured by Mutual Information (MI) in the COCA Newspaper corpus, and bigrams with weaker associations as measured by Approximate Collexeme (AC) in the COCA Spoken corpus.[3] Syntactic sophistication and complexity were also predictive of writing scores, such that higher-rated essays tended to have lower type-token ratios of verb-argument constructions found in the COCA academic corpus (i.e., greater repetitions of verb-argument constructions), and greater standard deviations of dependents per nominal complement (i.e., greater variety in the number of dependents per noun or noun phrase that follows copular verbs, such as “to be” and “to become”). Grammatical accuracy was also predictive of writing scores, such that fewer grammatical errors were an indication of higher-quality essays. Lastly, cohesion-related features predicted writing scores, such that higher-rated essays tended to have greater overlap of nouns and pronouns that occurred at least once in the next two paragraphs, and lower type-token ratio of adverbs (i.e., greater repetitions of adverbs).

[3] Both Mutual Information (MI) and Approximate Collexeme (AC) assess association strength of *n*-grams. MI represents the probability of co-occurrence of *n*-grams, being calculated as the logarithm of the observed co-occurrence of two lexical items divided by the expected co-occurrence of the two lexical items. *N*-grams with higher Mutual Information scores are those made up of strongly associated low-frequency words (e.g., *exultant triumph*; Evert, 2008). AC is calculated using the negative log of Fisher-Yates exact test (Fisher, 1925; Yates, 1934).

**Table 2**

*Results of the LME Model Predicting Writing Scores*

| Fixed effect | Estimate | Standard error | *t* | *p* |
|---|---|---|---|---|
| (Intercept) | 5.175 | 1.077 | 4.805 | < .001 |
| Differential Word Use (DWU) | .971 | .058 | 16.795 | < .001 |
| Adjacent Two–Paragraph Overlap of Noun and Pronoun | .803 | .167 | 4.817 | < .001 |
| Age of Exposure (Inverse Slope) | 2.008 | .521 | 3.851 | < .001 |
| COCA Fiction Frequency Function Word Type | .000 | .000 | –3.760 | < .001 |
| COCA Academic Frequency Content Word Type (Logarithm) | –1.076 | .288 | –3.732 | < .001 |
| COCA Newspaper Bigram Association Strength (MI) | –1.118 | .360 | –3.101 | < .010 |
| Adverb Type–Token Ratio | –.698 | .226 | –3.092 | < .010 |
| COCA Academic Construction Type–Token Ratio | .474 | .191 | 2.487 | < .050 |
| Standard Deviation of Dependents Per Nominal Complement | .137 | .060 | 2.296 | < .050 |
| COCA Spoken Bigram Association Strength (AC) | –.001 | .001 | –2.211 | < .050 |
| Number of Grammatical Errors (Normed for essay length) | –16.841 | 7.875 | –2.139 | < .050 |
| Time (Time 1 baseline) | .002 | .093 | .018 | .985 |
| Prompt (Originality baseline) | –.002 | .101 | –.015 | .988 |
| Time x Prompt | –.139 | .145 | –.954 | .342 |
| Class level: Advanced Composition (Composition I baseline) | –.081 | .114 | –.714 | .503 |
| Class level: Composition II (Composition I baseline) | –.046 | .154 | –.300 | .776 |
| Class level: Advanced Composition (Composition II baseline) | –.035 | .121 | –.290 | .779 |

### 5.3 Differences Over Time in Student Production of Linguistic Features

We examined whether there were changes over time in the 11 linguistic features that significantly predicted holistic writing scores in the final LME model.  Paired *t* tests were conducted with the alpha level adjusted to .005 using a Bonferroni correction for multiple comparison (see Table 3). Results of the *t* tests indicated that three linguistic variables showed significant differences over time and small effect size (Cohen's *d* > .200): COCA Fiction Frequency Function Word Type, COCA Academic Construction Type-Token Ratio, and COCA Academic Frequency Content Word Type. Two additional variables showed at least a small effect size (although not significant with Bonferroni correction): COCA Spoken Bigram Association Strength (AC) and Adjacent Two-Paragraph Overlap of Noun and Pronoun. These results indicate that over time, students tended to use more function words that occurred less frequently in a fiction corpus and content words that occurred less frequently in an academic corpus. In addition, writing over time contained greater repetitions of verb-argument constructions that are frequently used in an academic corpus. Also, student essays over time tended to include fewer bigrams that were strongly associated with speech. Finally, writing over time included greater overlap of nouns and pronouns across adjacent paragraphs. Importantly, changes in all of these five indices over time match the directions to which each index predicted writing quality. If an index negatively predicted writing score, its average score decreased over time. On the contrary, if an index positively predicted writing score, its average score increased over time.

**Table 3**

*Results for Paired T Tests for Linguistic Features*

| Index | *M* (*SD*) at Time 1 | M (SD) at Time 2 | *t* | *p* | Cohen's *d* |
|---|---|---|---|---|---|
| COCA Fiction Frequency Function Word Type | 8505.425 (1202.548) | 8052.447 (1133.626) | 5.891 | .000 | .373 |
| COCA Academic Construction Type-Token Ratio | .767 (.084) | .741 (.082) | 4.109 | .000 | .262 |
| COCA Academic Frequency Content Word Type (Logarithm) | 2.296 (.127) | 2.263 (.122) | 3.489 | .001 | .220 |
| COCA Spoken Bigram Association Strength (AC) | 15352.790 (3665.271) | 14659.290 (3124.626) | 2.442 | .015 | .204 |
| Adjacent Two-Paragraph Overlap of Noun and Pronoun | .330 (.187) | .363 (.164) | –2.448 | .015 | .206 |
| Differential Word Use (DWU) | –.120 (.522) | –.029 (.525) | –2.257 | .025 | .174 |
| Grammatical Error Count (Normed for essay length) | .002 (.004) | .002 (.003) | 1.799 | .073 | N/A |
| Adverb Type-Token Ratio | .640 (.133) | .621 (.132) | 1.786 | .075 | .143 |
| Standard Deviation of Dependents Per Nominal Complement | .884 (.527) | .912 (.473) | –.701 | .484 | .056 |
| Age of Exposure (Inverse Slope) | 1.032 (.067) | 1.036 (.061) | –.689 | .491 | N/A |
| COCA Newspaper Bigram Association Strength (MI) | 1.646 (.161) | 1.654 (.141) | –.611 | .541 | N/A |

## 6.0 Discussion

Understanding how linguistic features predict writing quality and how they change over time as writers advance in their first year of college can provide us with important indications of writing success and better explain how linguistic features interact with this success. The purpose of this article was to examine how individual linguistic features in students' writing were predictive of writing quality and whether these features changed as a function of time across a semester of composition study at a large university in the southeast of the United States. In addition, this study examined potential intervening variables that may explain writing quality, including course type, teacher, and prompt. Lastly, this study examined whether linguistic features demonstrated changes in students' writing as a result of time.

Our LME analysis revealed that a number of linguistic features were predictive of writing quality. However, the analysis indicated that time, class level, and prompt were not predictive of writing quality when co-varied with linguistic features. In terms of writing quality, 11 linguistic features were predictive. The strongest predictor of writing quality was differential word use, which is a rarely used metric that measures the frequency of words in the text that occur in low- and high-quality essays. Our DWU measure indicates that higher proficiency writers produce a greater number of words that are common in high-quality essays demonstrating greater knowledge of expected content. Our next highest predictor was a global cohesion feature that measures the number of pronouns and nouns shared across paragraphs. Essays showing greater global cohesion (i.e., more overlap across paragraphs) scored higher. The next three linguistic features were related to lexical sophistication. The first index measures the predicted age of exposure of the words produced. The findings indicate that higher-quality essays contained words that were predicted to be learned later. The next two lexical indices were related to word frequency and demonstrated that higher-quality writers used more infrequent words, both function words (taken from the COCA fiction sub-corpus) and content words (taken from the COCA academic sub-corpus).

The next two strongest predictors were related to word and structural diversity (i.e., the repetition of words and syntactic constructions). These predictors indicated that higher-quality writing samples included greater repetitions of adverbs and academic constructions (i.e., less diversity of adverbs and structures). Of the final four features, two were related to the association strengths between words. These two features indicated that higher-quality essays contained more word combinations that were strongly associated with academic writing and fewer word combinations that were strongly associated with speech, demonstrating higher scores for those students who had moved toward meeting academic writing expectations and away from spoken discourse norms. A measure of syntactic complexity was also included in the model. This measure demonstrated that higher-quality essays showed greater variation in phrasal complexity such that more proficient writers showed a greater variation in the number of dependents per nominal complement compared to less proficient writers. The final linguistic feature was related to grammatical error counts and showed that more proficient writers made fewer grammatical errors.

The linguistic features that were predictive of essay quality in this study have been indicative of writing quality in previous studies that examined different writing prompts, writer populations, and writing tasks (including the writing tasks the students in this study completed in their composition classes). For instance, content knowledge on the part of the student or displayed within a text is generally accepted as an important indicator of writing quality across genres and tasks (Attali, 2011; Gradwohl & Schumacher, 1989; Hayes, 1996, 2011). As well, increased global cohesion, generally across paragraphs, is predictive of more proficient writing samples (Crossley, Roscoe, et al., 2011; Crossley & McNamara, 2011), as is more advanced lexical production (Crossley, Weston, et al., 2011; Douglas, 2013; McNamara et al., 2010). Likewise, previous research has indicated that more proficient writers produce more target-like word combinations and more strongly associated combinations (Granger & Bestgen, 2014; Paquot, 2019) as well as more complex syntactic structures (Benson & Campbell, 2009; Klecan-Aker & Hendrick, 1985; Myhill, 2008). Lastly, a number of studies have indicated that a greater number of grammatical and mechanical errors is associated with lower writing proficiency (Crossley et al., 2014; Crossley et al., 2019; Morris et al., 1995).

One large question we addressed in our writing quality analysis was whether time (pretest or posttest) had an effect on essay quality. We found that writers in this study did not show gains between pretest and posttest essays when co-varied with linguistic features. While there were descriptive gains (see Table 1), this time variable was not a significant predictor of essay quality when considered in tandem with linguistic features. As well, no interactions were reported between class and time (i.e., no differences in times were noted based on class levels) in predicting essay quality when linguistic features were included in the model. The main reason for this finding is likely that linguistic features are much stronger predictors of essay quality than time and the linguistic features explain similar variance in essay quality as time. To explore if essay quality did change as a function of time, we ran a post-hoc pairwise *t* test to examine differences between pretest and posttest essay scores. The *t* test ($t = –2.812$, $p < .010$) indicated that there were significant differences between pretest and posttest scores for the student writing sampled such that posttest essays were scored higher.

In addition to time, we saw no effects for predicting text quality in terms of class or prompt when linguistic features were included in the model. For class, while there were differences in the scores between the class levels descriptively (see Table 1), these differences were not significant predictors of essay quality when linguistic features were co-varied, likely because linguistic features are stronger predictors of essay quality than class level. A post-hoc analysis did demonstrate difference in writing quality between class levels, $F(2, 497) = 13.027$, $p < .001$. Tukey's multiple comparison of means indicated that significant differences were reported between Advanced Composition class essay scores and Composition I ($p < .001$) and Composition II ($p < .050$) essay scores such that Advanced Composition class essay scores were stronger. It should also be noted that the types of writing required in the composition classes did not always match the standardized writing assessments used for data in this study. Much of the writing in the composition classes was not timed, and most of the writing was synthesis or source-based writing that required students to integrate information from external texts. Of the five required writings, three were argumentative, while two focused on literary analyses, which could be argumentative, but argumentation was not required. Of the three argumentative writings, only one was timed (the final exam), and the writing task was to integrate information from four sources into an argumentative essay. The remaining two writing assignments were not timed and were integrated writing tasks. Thus, lack of differences in the full models between composition classes may reflect the simple notion that the students did not practice the same writing tasks in class as they did on the pre and posttest.

We also found no effects for prompt on predicting essay score when linguistic features were co-varied. This may indicate that the two prompts selected for this analysis were generally equivalent in terms of leading to essays of similar quality. However, post-hoc paired *t* tests demonstrate that this is not the case ($t = –2.200$, $p < .050$), with the Uniqueness prompt ($M = 3.410$, $SD = .978$) leading to higher essay scores than the Heroes prompt ($M = 3.222$, $SD = .973$). Thus, there were differences in text quality based on the prompts, but these differences were not strong predictors of essay quality when co-varied with linguistic features.

Overall, we found that linguistic features were the strongest predictors of essay quality and explained a significant amount of variance (~60% of the score) even when time, class level, and prompt were taken into consideration. The strongest predictor of essay quality was a measure of content (DWU) followed by a measure of global cohesion, measures of lexical sophistication, word and structure diversity measures, a syntactic complexity feature, and features related to word association strength and grammatical accuracy. In terms of predicting essay quality, when co-varied with linguistic features, time, class level, and prompt were not significant predictors even though they demonstrated differences in univariate analyses. This finding helps to support the notion that linguistic features are the strongest predictors of essay quality and provides evidence that various linguistic features work in tandem (Crossley et al., 2015; MacArthur et al., 2019; McNamara et al., 2010). It is also interesting to note that random variance at the student and teacher level only explained a small amount of the variance (~3%) when co-varied with linguistic features.

Our second research question asked if linguistic features changed over time as a function of completing a university-level composition course. We limited the number of features we examined to those that were significant predictors of essay quality so that we could assess if longitudinal differences existed for meaningful variables. We found that of the 11 variables, five of them demonstrated at least small effect sizes (Cohen's $d > .200$). However, only three of these were considered significant because we lowered our alpha value to control for multiple comparisons. Of the variables that showed at least a small effect size, four of them were related to COCA indices that measured word frequency, construction TTR, and bigram association strength. The results indicated that over time, students produced less frequent words (both function and content words), greater repetitions of verb-argument constructions found in an academic corpus, and bigrams that had lower association strength found in a spoken corpus, indicating that students gradually produce more academic language in terms of words and structures. The remaining variable was related to global cohesion and demonstrated that over time, students produced essays with increased noun and pronoun overlap between paragraphs, indicating greater global cohesion. There were no changes in DWU scores, grammatical accuracy, adverb TTR, or variation in dependents in nominal complements. In total, the findings indicate that over the course of the semester, students used more academic language and developed better discourse organization skills (i.e., increased global cohesion), generally supporting previous studies (Bereiter & Scardamalia, 1987; Berman & Verhoeven, 2002; Crossley & McNamara., 2011; Flower & Hayes, 1980). However, little support was found for increased levels of syntactic complexity over time, which is similar to that reported by MacArthur et al. (2019) but different from other studies (Haswell, 2000; Myhill, 2009).

The intersection of the linguistic features that predict writing quality and the linguistic features that developed longitudinally provides some evidence for teaching approaches in the writing classroom designed to practice the inclusion of linguistic features in writing. We say this even given the differences between the writing tasks in the pre and posttest design and the tasks found in the classroom because there seem to be a small number of linguistic items that are both predictive of writing quality in an independent writing task and show propensity for growth in a short time frame in which students attended composition classes that focus on integrated writing tasks. These features, which include word frequency, construction TTR, bigram association strength, and global cohesion, should be considered as potential topics of instruction in lower-level composition courses because they demonstrate trends in a range of writing tasks and therefore may be generalizable across tasks. Specifically, students could be introduced to more sophisticated words (i.e., less frequent) that also have strong associations with other words to help writers not only use more complex words, but words that make meaningful associations with other words. This would require the use of frequency lists that include association metrics, similar to that found in the Academic Formula Lists (Simpson-Vlach & Ellis, 2010). In terms of global cohesion, students could be instructed to ensure that ideas, specifically nouns, are shared across paragraphs to increase cohesion throughout the text.

The methods used in this study, along with the results, also have implications for writing analytics and automated evaluation of writing quality and development. From a writing analytics perspective, the use of NLP tools to examine not only writing quality but also writing development simultaneously is a unique addition to available techniques. Since the tools used in this study are open source, the approaches used here could strengthen future writing analytic studies exploring links between linguistic features and writing development and quality. Additionally, the results have implications for writing analytics in practice, specifically for automated essay scoring (AES) and automated writing evaluation (AWE) systems. The linguistic features used in this study could be included in AES systems to provide students with summative feedback in the form of holistic essay scores and formative feedback in terms of global cohesion, strength of word associations, and vocabulary choices. Feedback on these features may be especially pertinent because they relate to writing development during the first year of matriculation in a university setting.

It is also important to highlight limitations of this study in terms of the writing type it analyzed, specifically standardized academic writing. While timed, independent writing is not uncommon in the writing classroom or in assessment, it is not representative of the types of writing found in the composition classroom or outside of academia, as evidenced by the actual writing reported in the composition courses in this study. All writing is situated, contingent, and genre- and task-specific, and the writing sample analyzed here focuses on a narrow, potentially secondary writing task. Additionally, we depended on an SAT scoring rubric to operationalize writing quality, and some research shows that SAT writing scores do not show strong internal consistency and alternate-form reliability estimates (Ewing et al., 2005). Thus, it is an open question whether the results reported in this study would generalize to other types of writing including, but not limited to, integrated writing, narrative writing, journal writing, expository writing, research reports, or other types of academic and professional writing. However, we take some reassurance that the writing development seen within the participants in this study overlapped with estimates of writing quality in the standardized writing task, potentially indicating universal tendencies between the development and assessment.

## 7.0 Conclusion

Understanding links between linguistic features in text and both writing quality and writing development can provide important information about the writing process and how writing skills change over time. This study examined pretest and posttest writing samples taken from 250 students across three different college-level composition classes and examined how linguistic features in the samples were predictive of essay quality when co-varying time, class level, and prompt. We found that essay quality was most strongly predicted by linguistic features related to content, lexical sophistication, global cohesion, syntactic complexity, and grammatical accuracy. We found that time, class level, and prompt were not significant predictors of essay quality when co-varied with linguistic features, although post-hoc analyses did find that essay scores differed among pretests and posttests, class levels, and prompts.

In addition, this study analyzed changes in linguistic features that were predictive of essay quality from pretest and posttest to examine longitudinal changes in linguistic production. We found that variables related to lexical sophistication, syntactic constructions, and global cohesion demonstrated differences indicating that students' writing changed as a function of time. We also found that the changes followed trends in the writing quality analysis such that the students began to produce linguistic patterns that should lead to increased essay quality scores. This finding provides evidence that over time, college-level students began to make linguistic changes in their writing.

This study provides additional evidence for relationships between linguistic features and essay quality as well as longitudinal growth in linguistic production. It builds on a series of previous studies that have shown similar trends in writing quality, but also adds additional levels of detail and a greater focus on longitudinal growth, where research is rare. It calls for the need for additional longitudinal studies, especially those that are mixed with essay quality analyses, to build on the current findings. Specifically, the need exists for larger longitudinal studies across a greater variety of grade levels and ages and across a greater range of writing tasks. Additionally, longitudinal studies of greater length (beyond a semester of study) are needed, as are studies that better control for potential demographic and individual differences among writers.

## Author Biographies

**Scott A. Crossley** is a Professor of Applied Linguistics and Learning Sciences at Georgia State University. Professor Crossley's primary research focus is on natural language processing and the application of computational tools and machine learning algorithms in language learning, writing, and text comprehensibility. His main interest area is the development and use of natural language processing tools in assessing writing quality and text difficulty.

**Minkyung Kim** is an applied linguist working as an assistant professor at Nagoya University of Commerce and Business in Japan. Her research focuses on second language writing and language assessment and has appeared in *Assessing Writing*, *The Modern Language Journal*, and *Journal of Second Language Writing*.

## References

Applebee, A. N. (2002). Engaging students in the disciplines of English: What are effective schools doing? *The English Journal*, *91*(6), 30-36.

Attali, Y. (2011). A differential word use measure for content analysis in automated essay scoring. *ETS Research Report Series*, *2011*(2), 1-19.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445-459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bartoń, K. (2017). *MuMIn: Multi-Model Inference. R package version 1.40.0.* https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-9*. https://CRAN.R-project.org/package=lme4

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge University Press.

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

Benson, B. J., & Campbell, H. M. (2009). Assessment of student writing with curriculum-based measurement. In G. A. Troia (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices* (pp. 337-357). Guilford Press

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written communication*. Lawrence Erlbaum.

Berge, K. L., Skar, G. B., Matre, S., Solheim, R., Evensen, L. S., Otnes, H., & Thygesen, R. (2019). Introducing teachers to new semiotic tools for writing instruction and writing assessment: Consequences for students' writing proficiency. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 6-25.

Berman, R., & Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language and Literacy*, *5*(1), 1-43.

Berninger, V., Fuller, F., & Whitaker, D. (1996). A process approach to writing development across the life span. *Educational Psychology Review*, *8*, 193-218.

Berninger, V., Nagy, W., & Beers, S. (2011). Child writers' construction and reconstruction of single sentences and construction of multi-sentence texts: Contributions of syntax and transcription to translation. *Reading and Writing. An Interdisciplinary Journal*, *102*, 151-182.

Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing. An Interdisciplinary Journal*, *4*, 257-280.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*, 5-35.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series*, *2013*(2), i-15.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*, 45-50.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977-990.

Cameron, C. A., Lee, K., Webster, S., Munro, K., Hunt, A. K., & Linton, M. J. (1995). Text cohesion in children's narrative writing. *Applied Psycholinguistics*, *16*(3), 257-269.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234.

Cox, B. E., Shanahan, T., & Sulzby, E. (1990). Good and poor elementary readers' use of cohesion in writing. *Reading Research Quarterly*, *26*, 47-65.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251-270.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. M. McCarthy & G. M. Youngblood (Eds.). *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. (pp. 214-219). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Kyle, K., Allen, L., & McNamara, D. S. (2014). The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th Educational Data Mining (EDM) Conference* (pp. 300-303). Springer.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment*, *8(*1). https://escholarship.org/content/qt1f21q8ck/qt1f21q8ck.pdf?t=r071l4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227-1237.

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

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

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*(4), 66-79.

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, *7*(3), 351-370.

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 438-440). Springer.

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282-311.

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology*, *100*(4), 907-919.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159-190.

Douglas, R. D. (2013). The lexical breadth of undergraduate novice level writing competency. *The Canadian Journal of Applied Linguistics*, *16*(1), 152-170.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(02), 143-188.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17-44.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1212-1248). Mouton de Gruyter.

Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring of cohesive devices. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 177-181). Newbury House.

Ewing, M., Huff, K., Andrews, M., & King, K. (2005). Assessing the reliability of skills measured by the SAT®. *Research Notes*. RN-24. College Board.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, *31*(1), 21-32.

Fox, J., & Weisberg, S. (2010). *An R companion to applied regression*. SAGE Publications, Inc.

Gere, A. R. (2019). *Developing writers*. University of Michigan Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Gradwohl, J. M., & Schumacher, G. M. (1989). The relationship between content knowledge and topic choice in writing. *Written Communication, 6*(2), 181-195.

Graham, S. (1983). Effective spelling instruction. *Elementary School Journal*, *83*(5), 560-567.

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, *89*(1), 170-182.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics, 52*, 229–252. https://doi.org/10.1515/iral-2014-0011

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

Haswell, R. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, *17*(3), 307-352.

Haswell, R. H.  (1986). *Change in undergraduate and post-graduate writing performance: Quantified findings*. (ERIC Document Reproduction Service No. ED 269 780.)

Haswell, R. H. (1990). *Change in undergraduate and post-graduate writing (Part 2): Problems in interpretation*. ERIC Clearinghouse on Reading and Communication Skills, ED 323 537.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1–27). Mahwah, NJ: Lawrence Erlbaum Associates.

Hayes, J. (2011). Kinds of knowledge-telling: Modeling early writing development. *Journal of Writing Research, 3*, 73–92.

Hunt, K. (1965). *Grammatical structures written at three grade levels*. NCTE.

Hunt, K. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the Society for Research in Child Development*, *35*(1), 1-67.

Hunt, K. W. (1966). Recent measures in syntactic development. *Elementary English*, *43*, 732-739.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, *4*(1), 51-69.

Ishikawa, S. I. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, *1*, 91-118.

Jagaiah, T., Olinghouse, N. G., & Kearns, D. M. (2020). Syntactic complexity measures: Variation by genre, grade-level, students' writing abilities, and writing quality. *Reading and Writing*, *33*, 2577-2638.

King, M., & Rentel, V. (1979). Toward a theory of early writing development. *Research in the Teaching of English*, *13*, 243-253.

Klecan-Aker, J. S., & Hendrick, D. L. (1985). A study of the syntactic language skills of normal school-aged children. *Language, Speech, and Hearing Services in Schools*, *16*(3), 187-198.

Kress, G. (1994). *Learning to Write.* London: Routledge.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). *lmerTest: Tests in linear mixed effects models. R package version 2.0-29*. http://CRAN.R-project.org/package=lmerTest

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. ScholarWorks @ Georgia State University. https://scholarworks.gsu.edu/alesl_diss/35

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786.

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513-535.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*, 1030-1046.

Langacker, R. W. (2007). Cognitive grammar. In D. Geeraets & H. Cuyckens (Eds.), *The Oxford handbook of cognitive linguistics* (pp. 421-462). Oxford University Press. https://doi.org/10.1017/s0022226709005775

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307-322.

Loban, W. D. (1976). *Language development: Kindergarten through grade twelve* (Research Report Number 18). National Council of Teachers of English.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, *32*(6), 1553-1574.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85-104.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392.

McCulley, G. A. (1985). Writing quality, coherence, and cohesion. *Research in the Teaching of English*, *19*, 269-282.

McCutchen, D., & Perfetti, C. (1982). Coherence and connectedness in the development of discourse production. *Text*, *2*, 113-139.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57-86.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499-515.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1-43

Morris, D., Blanton, L., Blanton, W. E., & Perney, J. (1995). Spelling instruction and achievement in six classrooms. *The Elementary School Journal*, *96*(2), 145-162.

Myhill, D. (2008). Towards a linguistic model of sentence development in writing. *Language and Education*, *22*(5), 271-288.

Myhill, D. (2009). Becoming a designer: Trajectories of linguistic development. In R. Beard, D. Myhill, & M. Nystrand (Eds.), *The SAGE handbook of writing development* (pp. 402-414). SAGE Publications.

Neuner, J. L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English*, *21*, 92-105.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121-145. https://doi.org/10.1177/0267658317694221.

Perera, K. (1984). *Children's writing and reading: Analysing classroom language*. Blackwell.

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, *28*, 56-78

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sanders, T., & Pander Maat, H. (2006). Cohesion and coherence. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (Vol. 2, pp. 591-595). Elsevier.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, *22*(1), 69-90.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Siyanova–Chantura, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, *36*, 549-569.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598-605.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149-183.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education*, *131*, 33-48.

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing, 18*, 187-201.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's wordbook of 30,000 words*. Columbia University, Teachers College: Bureau of Publications.

Tremblay, A., & Ransijn, J. (2015). *LMERConvenienceFunctions: Model selection and post-hoc analysis for (G)LMER models*. *R package version 2.10*. https://CRAN.R-project.org/package=LMERConvenienceFunctions

Verhoeven, L., Aparici, M., Cahana-Amitay, D., van Hell, J., Kriz, S., & Viguié-Simon, A. (2002). Clause packaging in writing and speech: A cross-linguistic developmental analysis. *Written Language and Literacy*, *5*(2), 135-161.

Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Wilson, L. G., Tschinkel, E., Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing*, *24,* 203-220.

Wharton-McDonald, R., Pressley, M., & Hampton, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and student achievement. *The Elementary School Journal*, *99*(2), 101-128.

Witte, S., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication*, *32*, 189-204.

Yates, F. (1934). Contingency tables involving small numbers and the $\chi^2$ test. *Journal of the Royal Statistical Society*, *1*(2, Suppl), 217-235.

Zhu, W. (2004). Faculty views on the importance of writing, the nature of academic writing, and teaching and responding to writing in the disciplines. *Journal of Second Language Writing*, *13*(1), 29-48.

# Appendix A: Writing Prompts Used in the Corpus Collection

| Topic | Prompt | Assignment |
|---|---|---|
| Heroes | Having many admirers is one way to become a celebrity, but it is not the way to become a hero. Heroes are self-made. Yet in our daily lives we see no difference between “celebrities” and “heroes.” For this reason, we deprive ourselves of real role models. We should admire heroes—people who are famous because they are great—but not celebrities—people who simply seem great because they are famous. | Should we admire heroes but not celebrities?<br>Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations. |
| Uniqueness | We value uniqueness and originality, but it seems that everywhere we turn, we are surrounded by ideas and things that are copies or even copies of copies. Writers, artists, and musicians seek new ideas for paintings, books, songs, and movies, but many sadly realize, “It's been done.” The same is true for scientists, scholars, and businesspeople. Everyone wants to create something new, but at best we can hope only to repeat or imitate what has already been done. | Can people ever be truly original?<br>Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations. |

## Appendix B: Essay Scoring Rubric

After reading each essay and completing the analytical rating form, assign a holistic score based on the rubric below. For the following evaluations you will need to use a grading scale between 1 (minimum) and 6 (maximum). As with the analytical rating form, the distance between each grade (e.g., 1-2, 3-4, 4-5) should be considered equal.

SCORE OF 6: An essay in this category demonstrates clear and consistent mastery, although it may have a few minor errors. A typical essay effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position; is well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas; exhibits skillful use of language, using a varied, accurate, and apt vocabulary; demonstrates meaningful variety in sentence structure; and is free of most errors in grammar, usage, and mechanics.

SCORE OF 5: An essay in this category demonstrates reasonably consistent mastery, although it will have occasional errors or lapses in quality. A typical essay effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position; is well organized and focused, demonstrating coherence and progression of ideas; exhibits facility in the use of language, using appropriate vocabulary; demonstrates variety in sentence structure; and is generally free of most errors in grammar, usage, and mechanics.

SCORE OF 4: An essay in this category demonstrates adequate mastery, although it will have lapses in quality. A typical essay develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position; is generally organized and focused, demonstrating some coherence and progression of ideas; exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary; demonstrates some variety in sentence structure; and has some errors in grammar, usage, and mechanics.

SCORE OF 3: An essay in this category demonstrates developing mastery, and is marked by ONE OR MORE of the following weaknesses: develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position; is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas; displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice; lacks variety or demonstrates problems in sentence structure; and contains an accumulation of errors in grammar, usage, and mechanics.

SCORE OF 2: An essay in this category demonstrates little mastery, and is flawed by ONE OR MORE of the following weaknesses: develops a point of view on the issue that is vague or seriously limited; demonstrates weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position; is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas; displays very little facility in the use of language, using very limited vocabulary or incorrect word choice; demonstrates frequent problems in sentence structure; and contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured.

SCORE OF 1: An essay in this category demonstrates very little or no mastery, and is severely flawed by ONE OR MORE of the following weaknesses: develops no viable point of view on the issue, or provides little or no evidence to support its position; is disorganized or unfocused, resulting in a disjointed or incoherent essay; displays fundamental errors in vocabulary; demonstrates severe flaws in sentence structure; and contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning.

Holistic score based on attached rubric (1-6): ___