Linguistic features in writing quality and development: An overview

Scott Crossley

Georgia State University, Atlanta (GA) | USA

Abstract: This paper provides an overview of how analyses of linguistic features in writing samples provide a greater understanding of predictions of both text quality and writer development and links between language features within texts. Specifically, this paper provides an overview of how language features found in text can predict human judgements of writing proficiency and changes in writing levels in both cross-sectional and longitudinal studies. The goal is to provide a better understanding of how language features in text produced by writers may influence writing quality and growth. The overview will focus on three main linguistic construct (lexical sophistication, syntactic complexity, and text cohesion) and their interactions with quality and growth in general. The paper will also problematize previous research in terms of context, individual differences, and reproducibility.

Keywords: Cohesion, Lexical sophistication, Syntactic complexity, Writing quality, Writing development, Linguistics

journal of WRITING RESEARCH

Crossley, S.A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3), 415-443. doi: xx

Contact:Scott A. Crossley, Department of Applied Linguistics/ESL, Georgia State University, 25 Park Place, Suite 1500, Atlanta, GA 30303 | USA - sacrossley@gmail.com

Linguists and writing researchers have long sought ways to examine how language features in texts relate to both writing quality and writing development in both first language (L1) and second language (L2) writers. Early analyses were conducted on small samples of texts using hand-coded features. While these studies provided important information about how linguistic features in the text interacted with quality and growth, they were impractical, difficult to scale up, and prone to mistakes. Recent innovations in natural language processing (NLP) have provided means with which to better calculate linguistic features in large corpora of writing samples which has afforded insights into a number of cognitive phenomena including human judgements of text quality and longitudinal writing growth. These innovations have spearheaded research into not only better understanding the linguistic elements of text that help predict writing quality and development, but also research into automatic essay scoring (AES) systems that provide summative feedback to writers about overall writing quality and automatic writing evaluation (AWE) systems that provide formative feedback to assist writers in revising specific aspects of writing. The robustness of such systems and their practicality have led to the development of a number of commercial application to capitalize on the success of AES and ASE systems, providing real-time feedback to student writers, and helping classroom teachers and administrators manage resources better.[1]

The purpose of this paper is to provide an overview of how analyses of linguistic features in L1 and L2 writing samples have afforded a greater understanding of predictions of both text quality and writer development.[2] Specifically, this paper provides an overview of how language features found in text can predict human judgements of writing proficiency and changes in writing levels in both cross-sectional and longitudinal studies. The goal is to provide a better understanding of how language features in text produced by writers may reflect writing quality and growth. Thus, the focus is on the linguistic features in text and not, per say, prediction accuracy of these features, although discussion of variance explained by linguistic features will be discussed later. Relatedly, the paper will not consider text length as a linguistic feature while acknowledging that text length is likely the strongest predictor of writing development and quality. Additionally, black box approaches such as neural network scoring algorithms that involve linguistic features but do not provide information about how these features interact with text quality will not be included. The overview will focus on three main linguistic construct (lexical sophistication, syntactic complexity, and text cohesion) and their interactions with writing quality and growth as well as problematize previous research in terms of writing context, individual differences, and reproducibility.

1. Linguistic Features and Writing Knowledge

Studies have linked specific linguistic elements of the written text to writing proficiency and development since the 1970s. The simple notion that the words writers produce

and the structure those words are placed along with larger coherence patterns are predictors of writing quality and development has been established in 100s of studies. Generally, a greater number of linguistic studies of writing quality come from the field of L2 studies because more L2 researchers are linguists whereas fewer writing researchers who focus on L1 writing are trained linguists. In addition, L2 writing is often used as a proxy of language ability in both standardized tests and language acquisition studies, leading to a greater interest in examining textual features.

In general, the linguistic features examined by writing researchers fall into three large constructs: lexical, syntactic, and cohesion (McNamara et al., 2010). Language features are also important elements of identifying discourse structures as well (i.e., claims, arguments, theses, and rhetorical moves), but the language structures used in these determinations may not be linguistic in nature per se and are not the focus of this overview. Instead, this review will focus on research that has examined links between writing quality/development and linguistic features as found in text, especially those features related to lexical sophistication, syntactic complexity, and text cohesion. I will treat these constructs separately for the purposes of providing an overview, but these features strongly interact with each other in terms of explaining writing quality and development. Additionally, it is well documented that linguistic features in writing vary based on individual differences (e.g., between L1 and L2 writers) but also on differences in writing prompts, topics, writing tasks, and discipline expectations. These differences will be covered in the discussion section of this paper.

I focus on writing quality because interest in writing assessment in terms of standardized and classroom testing as well as providing feedback to writers in intelligent tutoring system has led to much research into exploring the linguistic predictors of writing quality. While these approaches can tell us much about how linguistic features can distinguish between good and bad writers, they do not strongly explain development. Thus, I also focus on writing development by reviewing a smaller number of writing studies have examined differences in linguistic output based on grade level or longitudinal studies. Both of these approaches give us stronger indications of how linguistic features develop over time in writers and a better understanding of learning progressions (i.e., developmental sequence that demonstrates a vertical continuum of increasing expertise over time, Masters & Forster, 1996; Popham, 2007; Wilson & Bertenthal, 2005).[3]

2. Lexical Sophistication

Lexical items are perhaps the most commonly used linguistic feature to analyze the quality of texts. Quality of lexical items can be subsumed under the term lexical richness which generally consists of lexical diversity (i.e., the number of unique words), lexical density (i.e., the number of content to function words), and lexical sophistication. Lexical sophistication tends to provide the richest metrics of text quality and can be thought of as the proportion of advanced words in a text (Read, 2000).

Traditionally, sophisticated words have been operationalized as low frequency words (Laufer & Nation, 1995), but this has changed over time such that sophistication can encompass a vast number of word properties. For instance, sophisticated words have been defined as words that are more likely found in academic text (Coxhead, 2000), words that are less concrete, imageable, and familiar (Crossley & Skalicky, in press; Salsbury, Crossley, & McNamara, 2011; Saito et al., 2016), words that have fewer phonological and orthographical neighbors, words that have higher latencies in word naming and lexical decision tasks (Balota et al., 2007), more specific words (Fellbaum, 1998), and words that are less diverse based on context (McDonald & Shillcock, 2001). More recent research is pushing lexical sophistication away from single word properties and moving more toward multiword units under the presumption that two or more words combinations provide important indicators of lexical knowledge (Sinclair, 1991).

When assessing text quality, the existence of more sophisticated words in a writing sample is indicative of greater lexical knowledge and thus greater writing ability. There are a number of theoretical underpinnings for this. For instance, usage-based approaches to understanding lexical knowledge argue that elements such as frequency of occurrence, associative learning (i.e., establishing connections among words), automatization (i.e., producing words with less effort), abstraction (i.e., categorizing words into schemas), and developing representations of word form and meaning (Ellis, 2002; Langacker, 2007) all lead to lexical acquisition at the word and phrase level (Goldberg, 2006). From a psycholinguistic perspective, the properties of the words themselves influences recognition and processing (Balota et al., 2007). Multiple studies have demonstrated that words that elicit greater response times and are less likely to recognized as words are more sophisticated (i.e., less concrete and frequent). Psycholinguistic research has also found that knowledge of multi-word sequences gives users significant processing advantages (Ellis, 2012; Siyanova-Chantura & Martinez, 2015). From both perspectives, it is apparent that more proficient writers produce words that are more difficult to process and recognize either because of exposure to the words or because of properties inherent to the words.

2.1 Text quality and lexical sophistication

Lexical properties are strongly indicative of L1 and L2 writing quality, although more research has been reported for L2 writing for reasons stated before. In terms of L1 writing, research indicates that use of more academic words (Douglas, 2013), more specific words, more imageable words and less meaningful words (McNamara et al., 2013), longer words and less familiar words (Crossley, Weston, McLain, & McNamara, 2011), and a greater use of infrequent words (McNamara, Crossley, & McCarthy, 2010) is indicative of higher quality academic writing. At least one study also examined the sophistication of phrases in L1 writing (Crossley, Cai, & McNamara, 2012). This study reported that a number of phrasal features related to phrasal frequency and proportion (i.e., the number of phrases common in a reference corpus) were negatively correlated

with writing quality indicating that L1 writers who produces more sophisticated phrasal items were judged to be better writers.

Similar patterns have been reported for L2 writers such that higher quality texts are represented by more sophisticated words while writers develop over time to produce more sophisticated lexical items. For instance, like L1 studies, word frequency is predictive of human ratings of writing proficiency such that more proficient L2 learners tend to produce less frequent words, familiar, and meaningful words (Crossley & McNamara, 2012) and words with more letters or syllables (Grant & Ginther, 2000; Reppen, 1994). More proficient L2 writers also tend to use more specific words (Guo, Crossley, & McNamara, 2013; Kyle & Crossley, 2016) and less imageable words (Crossley, Kyle, Allen, Guo, & McNamara, 2014) than less proficient L2 writers.

Unlike L1 writing studies, a number of L2 studies have focused on phrasal sophistication in predicting writing quality. In general, these studies report that more proficient L2 writers produce a greater range of phrasal structure common in L1 language speech and writing samples and produce these structures more frequently than lower proficiency writers (Kyle & Crossley, 2015; Ohlrogge, 2009; Vidakovic & Barker, 2010). For instance, studies have shown that more proficient L2 writers produce more target-like bigrams and a greater number of strongly associated bigrams (Granger & Bestgen, 2014; Paquot, 2017). Kyle and Crossley (2015) also reported that more proficient L2 writers produced more frequent trigrams as found in the written portion of the BNC than lower proficiency writers. More recent studies have demonstrated that both bigram and trigram features related to proportion and association scores account for significant variance in human judgments of writing quality for Korean L2 writers of English (Garner, Crossley, & Kyle, 2018).

2.2 Writing development and lexical sophistication

A number of cross-sectional and longitudinal studies have examined how lexical features develop in L1 and L2 writers with results indicating that lexical features are also strong indicators of L1 writing development. In an early longitudinal study, Haswell (2000) that the number of words that were greater than nine letters increased over time in college level writing. In a more recent longitudinal study of basic college writers, MacArthur, Jennings, and Philippakos (2019) found that a lexical complexity component score increased in post-test essays written by basic college writing students. In a cross-sectional study, Crossley et al. (2011) examined differences between 9th-grade, 11th-grade, and college level essays and reported that the strongest discriminator of grade level was word frequency with college level writers producing more infrequent words. Other lexical properties of writing that changed across grade level included word concreteness and word polysemy. For concreteness, advanced writers started to produce more concrete terms (perhaps with respect to developing better claims) while for polysemy, the advanced writers began to produce words with fewer senses. A more recent study by Gardner, Nesi, and Biber (2019) using the British Academic Written

English (BAWE) corpus examined differences among three levels of undergraduate writing and one level of graduate writing. Gardner et al. used the Biber tagger (Biber, Johansson, Leech, Conrad, & Finegan, 1999) to tag the texts in BAWE for 150 different linguistic features and a multi-dimensional analysis was used to examine differences in linguistic features between discipline, level, and genre. The fourth dimension strongly distinguished between levels of writer and was informed by lexical items including long words, nominalizations, attributive adjectives, and abstract nouns, all of which are related to lexical sophistication. The analysis indicated that as student level increased, the number of lexical sophisticated words produced grew. The first dimension also discriminated texts by level, but to a lesser degree. The first dimension was informed by nouns as premodifiers, common nouns, concrete nouns, and quantity nouns, all of which showed a greater incidence in graduate writing as compared to undergraduate writing and level three undergraduate writing as compared to level 1 and 2 undergraduate writing.

Studies have also examined the development of lexical features in L2 writing, with many examining phrasal development. In an early study, Li & Schmitt (2009) found that L2 college writers in China produced a greater variety of lexical phrases over time. Cross-sectional studies have reported similar findings such as advanced L2 writers tending to produce more collocations than beginner and intermediate learners (Laufer & Waldman, 2011), more experienced EFL writers producing a greater range of frequent three-word lexical bundles (Leńko-Szymańska, 2014), and advanced college writers producing a greater number of phrases than beginning level college writers (Huang, 2015).

It should be noted that L2 writers rely on a smaller number of phrases, do not produce as many native-like sequences as L1 writers (Chen & Baker, 2010; Durrant & Schmitt, 2009), and overuse phrases common in L1 speech while underusing academically appropriate n-grams (Chen & Baker, 2010; Jukneviciene, 2009). Also, unlike lexical features, phrasal features in L2 writing do not develop to be more sophisticated, but rather develop toward being more acceptable (i.e., advanced L2 writers more strongly follow the patterns of L1 writers in terms of the proportion of phrases they use and the associations between the words in those phrases).

3. Syntactic Complexity

Another common approach to assess the quality of written text is to examine the syntactic properties of writing. Syntactic complexity refers to the sophistication of syntactic forms as well as the variety of syntactic forms produced (Lu, 2011; Ortega, 2003). The underlying notion is that more complex syntactic structures can act as an indicator of more advanced writing skills. Traditional approaches to measuring syntactic complexity have involved calculating sentence length with the notion that longer sentences are more complex and T-unit counts where a T-unit is a dominant clause and all subordinate clauses. Sentence length and T-unit calculations are consider

large-grain, length features (Kyle & Crossley, 2018) and have been the most common approaches to measuring syntactic

3.1 Syntactic Properties of Text Quality

Traditionally, syntactic complexity has been examined through large-grained, length-based syntactic indices known as T-units (Hunt, 1965). A T-unit is the shortest allowable grammatical unit punctuated at the sentence level. Thus, a T-unit can consist of a main clause plus additional, embedded subordinated clauses, but not two independent clauses joined together. T-units were used in early studies to examine L1 writing development (Hunt, 1965) and were later extended to L2 research for the same purpose (Lu, 2011; Ortega, 2003). Arguably, the use of T-units for writing analyses is more common in L2 writing. For example, in a synthesis of L2 writing studies, Ortega (2003) reported that over 90% of previous studies operationalized syntactic complexity as the mean length of T-unit. While common, T-units features are problematic because they often report conflicting results across studies (Bardovi-Harlig, 1992; Ortega, 2003; Stockwell & Harrington, 2003) and can be difficult to interpret (Norris & Ortega, 2009).

A good example of the interpretation problem can be illustrated through the feature mean length of T-units. While the mean length of the T-unit gives a general overview of amount of elaboration attached to a main clause, it provides no indication about the how, exactly, the clause is elaborated, which makes it difficult to calculate syntactic complexity in a fine-grained manner. As noted in Kyle and Crossley (2018), the two sentences

1a. The athletic man in the jersey kicked the ball over the fence.

1b. Because he wanted to score a goal, the man kicked the ball.

would both return a mean length of T-unit count of 12. However, the complexity in the first example rests on phrasal elaboration while the complexity in the second sentence rests on clausal elaboration. As noted by Biber, Gray, & Poonpon (2011), phrasal complexity is more strongly characteristic of academic writing while clausal complexity is characteristic of speech. However, using large-grained indices of clausal complexity like T-units would not distinguish between these two different types of complexity.

In response, a number of researchers have developed indices that measure more fine-grained indices of syntactic complexity including some T-unit indices including the number of clauses per T-unit and dependent clauses per clause, both of which measure clausal subordination (Ortega, 2003; Lu, 2011). Syntactic complexity indices have also been developed that are not based on T-units. For instance, the Biber tagger (Biber, 1988) tags a number of text features related to syntactic complexity including agentless passives, by-passives, that-verb and that-adjective complements, incidence of infinitives, phrasal and independent clause coordination, and a number of relative clause features. The Coh-Metrix tool (Graesser, McNamara, Louwerse, & Cai, 2004) includes a number of syntactic complexity features including indices related to the number of constituents in a sentence, the number of words before the main verb, and

syntactic similarity among sentences. Finally, the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC, Kyle, 2016) measures phrasal and clausal complexity features along with features related to the frequency profiles of verb argument constructions.

3.2 Text quality and syntactic complexity

Like lexical properties, syntactic features are indicative of both L1 and L2 writing quality. Also, like lexical features, the majority of syntactic complexity research has focused on L2 writers. In terms of L1 text quality, for children specifically, research has demonstrated that greater syntactic complexity in written texts equates to better increased writing scores (Benson & Campbell, 2009; Klecan-Aker & Hendrick, 1985). For instance, Myhill (2008) reported that better writers used fewer finite verbs, fewer finite subordinate clauses, and fewer coordinated clauses. Similar results have been found for L1 adult writers. For instance, McNamara et al. (2010) reported that better writers used greater syntactic complexity (i.e., a greater number of words before the main verb). Similar analyses indicated that the use of simple declarative sentences correlated negatively with essay quality while the length of noun phrases and the number of words before the main verb correlated positively (Crossley et al., 2011; McNamara et al., 2013). Not all studies report links between essay quality and syntactic complexity though. For instance, McNamara et al. (2013) and Perin and Lauterbach (2016) found no significant relationship.

Analysis of links between syntactic complexity features and L2 writing quality are more common than L1 writing quality, but the results are similar in that higher quality writing generally contains more complex syntactic features. A number of studies have operationalized syntactic complexity based on T-units (e.g. length of T-units, complex nominals per clause, and complex nominals per T-unit) and found that more proficient L2 writers produce longer and more varied syntactic structures (Lu, 2011; Ortega, 2003; Wolfe–Quintero, Inagaki, & Kim, 1998). Studies that go beyond T-units have indicated that higher rated L2 essays contain greater clausal subordination (Grant & Ginther, 2000) and incidence of passive structures (Connor, 1990; Ferris, 1994; Grant & Ginther, 2000). More recent studies indicate that higher quality L2 essays are defined by the production of dependent clause features such as the incidence of infinitives and 'that' verb complements (Crossley & McNamara, 2014) and a greater number of complex syntactic structures including syntactic structures related to clause complexity (*that* clauses and *to* clauses, Friginal & Weigle, 2014). Increased writing proficiency is also linked to greater phrasal complexity in writing (Taguchi, Crawford, & Wetzel, 2013). For instance, both Guo, Crossley, & McNamara (2013) and Jung, Crossley, & McNamara (in press) reported a positive relationship between mean length of noun phrases and writing quality while Kyle and Crossley (2018) reported that six indices phrasal features (related to nominal subject, direct object, and prepositional object modifiers) explained a significant amount of the variance in essay scores. In a direct

comparison between large-grain and fine-grain syntactic complexity indices, Kyle and Crossley (2017) examined differences between T-unit features and usage-based syntactic complexity features that measure lexicalgrammatical features (e.g., average main verb lemma frequency, verb-argument frequency) to predict L2 writing quality and found that fine-grained features predicted more than double the amount of variance in writing quality.

3.3 Writing development and syntactic complexity

Syntactic complexity features can also help explain writing development for L1 and L2 writers over time and across grade levels. For L1 writers, cross-sectional research has focused on complete sentence production while for adolescent and adults, the research shifts toward t-unit production and the incidence of syntactic embeddings and phrasal complexity. Research on the production of complete sentences reports that children show growth in their use of complete sentences over time while the production of run-on sentences and sentence fragments decreases (Berninger, Nagy, & Beers, 2011). T-unit research examining syntactic growth in young writers finds increased T-unit complexity as writers increase in grade level from 4th, 8th, and 12th grade (Hunt, 1965, 1966, 1970). These findings have been supported by Wagner et al. (2011) who found that 4th graders have a greater length of t-units and greater clause density than 1st graders. In terms of specific syntactic features that might inform increased T-unit length, research has found that writers produce a greater number of relative clauses, complement clauses, subordinate clauses, a wider variety of clause types, and a greater number of passives and modals as they advance in grade level (Berninger, Nagy, & Beers, 2011; Perera, 1984; Verhoeven et al., 2002) while coordinated clauses decreased (Verhoeven et al., 2002). Similar trends are reported for older students. For instance, Berman and Verhoeven (2002) found increases in mean length of clause between junior high school and high school writers while Crossley et al. (2011) found that college freshmen produced longer noun phrases than 11th graders and 9th graders. At the college level, Haswell (1986) reported that graduate students used more infinitives and had longer clause lengths than undergraduate students. In a second study, Haswell (1990) reported increased syntactic complexity for multi-clause unit spans (i.e., t-unit measures) and subclausal syntactic spans (i.e., clausal length especially at the noun phrase) between freshman and junior students and undergraduate and graduate students.

Longitudinal analyses of student writing development in terms of syntactic complexity have been rare, but do exist. As an example, Loban (1976) found that as children develop into adolescents and move from kindergarten to 12th grade, the produce longer sentences with more embedded clauses and longer noun phrases. For college writers, Haswell (2000) reported that writers, over time, begin to produce longer sentences with longer clauses, indicating syntactic growth. More recently,

MacArthur et al. (2019) analyzed pre- and post-test writings for basic college writers and found no differences in syntactic complexity features.

Syntactic development in L2 writers can also be assessed cross-sectionally and longitudinally. In an early cross-sectional study, Larsen-Freeman (1978) found that the percentage of error-free T-units and the average length of error-free T-units were strong predictors of L2 placement levels. Ferris (1994) reported that high level L2 writers produced more passives, nominalizations, relative clauses, adverbial clauses, and sentence conjuncts than lower level L2 writers (see Connor, 1990 for similar findings). In a synthesis report, Ortega (2003) found low and high proficiency L2 writers differed in their production of T-unit features including mean length of clause, mean length of T-unit, and clauses per T-unit. In a more recent study, Lu (2011) found a majority of T-unit features (10 of the 14 indices) showed differences between proficiency levels.

There have also been a number of longitudinal L2 studies that focus on syntactic complexity (Casanave, 1994; Ishikawa, 1995; Stockwell & Harrington, 2003; Crossley & McNamara, 2014). Casanave (1994) examined L2 syntactic growth in narrative writing over the course of three semesters and found L2 learners began to produce longer t-units, more error free t-units, and more complex t-units over time. Ishikawa (1995) examined low proficiency L2 writers and reported that syntactic accuracy features (error-free clauses per composition and total words in error-free clauses) showed differences between writings sampled at the beginning and end of semester. Difference over time have also been noted in short-term studies (five weeks) where L2 writers showed development over time in their average number of words per error-free T-unit, average number of words per T-unit, and percentage of error-free T-units. Byrnes (2009) reported that L2 German writers used more words per T-unit as a function of time spent studying German. Lastly, Crossley and McNamara (2014) reported significant growth in L2 writers' syntactic complexity as a function of time spent in a writing class. Specifically, L2 writers produced longer noun phrases, sentences that were less syntactic similar, more words before the main verb, and fewer verb phrases over the course of a semester. Similar findings have been reported for mean length of T-unit increases over a single semester of study (Bulté & Housen, 2014).

4. Text Cohesion

Text cohesion is related to the inter-connectivity of text segments of text based on textual features and is an important element of writing because it can indicate lexical, semantic, and argumentative dependencies within a text (Halliday & Hasan, 1976). Text cohesion can occur at the sentence level (i.e., local cohesion) or across larger segment gaps such as paragraph, chapters (i.e., global cohesion), or even texts (e.g., inter-document cohesion). Perhaps the most common approach to identifying cohesion within texts is to examine overt connections between text segments including referencing previous elements (generally through pronouns), repeating lexical items, substituting lexical items and the use of conjunctions to connect ideas. If cohesion in a

text is not maintained, it may be difficult for readers to evaluate the systematic relationship between shared lexical items, at which point a reader's mental representation of the text may break, affecting comprehension. The mental representation of a text that a reader develops is referred to as coherence (McNamara, Kintsch, Songer, & Kintsch, 1996; Sanders & Pander Maat, 2006) and the difference between cohesion and coherence is important. Cohesion is text-based and refers to the presence or absence of explicit cues in the text that afford connecting segments of texts together. Coherence, on the other hand, is reader-based and refers to the understanding that each individual reader or listener derives from the discourse While cohesion can be measured using text features, coherence can vary as a function not only of cohesion features but also individual differences in readers including background knowledge and language proficiency (McNamara et al., 1996).

4.1 Text quality and text cohesion.

Text cohesion has been an important element of assessing writing quality, especially for young L1 writers. In general, studies support the notion that local cohesion markers in young writers' texts are strong markers of quality (Englert & Hiebert, 1984; Struthers, Lapadat, & MacMillan, 2103). In terms of simple connectives and younger writers, writing samples judged to be lower quality are more likely to contain temporal adverbs while higher quality samples contain more causal, adversative, additive, and manner adverbials (Myhill, 2008). Additionally, Cox, Shanahan, and Sulzby (1990) found that the appropriate use of cohesive devices (co-referential devices like pronoun reference, ellipses, demonstratives) positively correlated with essay quality (for 3rd and 5th grade students) and Cameron et al. (1995) found that cohesion features (lexical cohesion, reference, and conjunction) accounted for a significant amount of variance in young students' writing quality.

For older writers (high school and college level writers), research on the use of text cohesive features is mixed, especially for local cohesion features. Early work by Witte and Faigley (1981) found greater density of cohesive ties in higher quality essays written by college students and more recent work has reported both significant negative (Perin & Lauterbach, 2016) and positive (MacArthur et al., 2019) relationship between referential cohesion (e.g., links between words across sentences) and text quality. However, research has reported no links in essay quality as a function of local cohesive devices including word and semantic overlap at the sentence level and the incidence of, positive logical connectives, logical operators, negative temporal connectives among many (Evola, Mamer, & Lentz, 1980; McCulley, 1985; McNamara et al., 2010; Neuner, 1987). Additional studies by Crossley and McNamara (2010; 2011) reported similar findings in that local cohesive devices including those reported in McNamara et al. (2010) in addition to causal, spatial, temporal cohesion features either do not correlate with human ratings of text coherence or correlate negatively to such ratings. In contrast, research does seem to indicate a clear link between global cohesive devices

and text quality though (Neuner, 1987). For instance, Crossley, Roscoe, McNamara, & Graesser (2011) found that two indices of global cohesion (semantic similarity between initial and middle paragraphs, and semantic similarity between initial and final paragraphs) significantly correlated with essay quality. Similar findings were reported by Crossley & McNamara (2011) and McNamara et al. (2013) who found that lexical and semantic overlap indices across paragraphs were positively correlated with ratings of essay quality. In addition, Crossley and McNamara (2016) found that student modifications to a text in terms of global cohesion led to increased quality scores.

Studies examining links between text quality and cohesion features in L2 writing are rarer than in L1 writing. Available studies demonstrate that adult L2 writers follow similar trends as L1 college level writers that more proficient L2 writers tend to produce less cohesive text as measured by lexical and semantic overlap across sentences (Crossley & McNamara, 2012; Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994). Similar results can be found in studies that focus on lexical diversity, which is related to referential cohesion in that lower lexical diversity signifies greater word overlap (McCarthy, 2005). These studies indicate that more proficient L2 writers produce texts with a greater diversity of words (i.e., less lexical overlap, Berman & Verhoeven, 2002; Crossley & McNamara, 2012; Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994). Associations between more explicit local cohesion features such as the use of connectives are less clear with some studies showing more proficient L2 writers producing more connectives (Jin, 2001; Connor, 1990), but at least one study (Crossley & McNamara, 2012) finding that higher quality L2 texts did not contain more connectives. A recent study college level L2 writers (Crossley & McNamara, 2016) reported that local cohesion features in general were negatively associated with essay quality (e.g., incidence of coordinating conjunctions and sentence overlap of pronouns) but function word overlap at the sentence level was predictive. Two global cohesion features (adjacent overlap between paragraphs for both function words and nouns) were positively related to essay quality.

4.2 Writing development and text cohesion

Studies examining how students develop in terms of text cohesion demonstrate movement from connecting ideas that local level and moving toward more global cohesion with time. For instance, L1 students initially connect ideas at the sentence level during writing (Berninger, Fuller & Whitaker 1996), but, with time, they start developing cohesion at the global level by linking topics across paragraphs (Bereiter & Scardamalia, 1987; Hayes & Flower 1980). At later stages, there may be movement away from the use of explicit cohesive devices in text and a movement toward the use of more complex syntactic structures to situate coherence (Haswell, 2000; McCutchen & Perfetti, 1982). All of this indicates that students in diverse grade levels use cohesive device differently (Crowhurst, 1987; Fitzgerald & Spiegel, 1986; Yde & Spoelders, 1985) In general, it appears that around the 2nd grade, students develop local cohesion

through the use of referential pronouns and connectives (King & Rentel, 1979) with a general increase in lexical repetition between 1st and 4th grades. Additionally, the distance between cohesion devices decreases with time such that referents become closer to one another (Fitzgerald & Spiegel, 1986; McCutchen & Perfetti, 1982; Yde & Spoelders, 1985). The development of local cohesion features continues until around the 8th grade (McCutchen & Perfetti, 1982) when student writings still contain more local cohesion devices than 6th grade writings, but students begin to use fewer explicit cohesion cues to organize text (McCutchen, 1986; McCutchen & Perfetti, 1982). In high school and college level writers, the development of more complex syntactic constructions can be seen (McCutchen & Perfetti, 1982) along with a decrease in the use of local cohesion devices (Crossley et al., 2011; Haswell, 1986). Difference also exist in variety as compared to number across grades with research showing that older students produce a greater variety, but not necessarily more, temporal and causal conjunctions (Crowhurst, 1987).

Fewer studies have focused on the development of cohesion devices in L2 writers. Yang and Sun (2012) examined differences between second and fourth-year undergraduate Chinese L2 English speakers and found that more advanced learners used a greater number of local cohesive devices (conjunctions, ellipsis, pronouns, and lexical overlap) and used them more accurately. In a longitudinal study, Crossley and McNamara (2016) found that college level L2 writers differed in pre- and post-test essays produced over the course of a semester. Specifically, they reported strong increases in noun overlap between paragraphs and in the semantic similarity between all sentences and paragraphs (i.e., global cohesion) as well as an increase in the repetition of content words and bigrams across a text (i.e., text cohesion). Weaker, but significant effects, were reported for semantic similarity between initial and final paragraphs, noun synonymy between paragraphs, and greater lexical overlap between sentences (all words and verbs)

5. Discussion

This paper presents a general overview of links between linguistic features in student writing and both writing quality and development. Overall, previous research has demonstrated clear and consistent associations between linguistic features and writing quality and development such that higher rated essays include more sophisticated lexical items, more complex syntactic features, and greater cohesion. Developing writers also show movements toward producing more sophisticated words and complex syntactic structures. Research also shows a movement away from the use of local cohesion devices in writing and a movement toward the development of more global cohesion features as a function of time with some research indicating that organizational flow may also begin to rely on more complex syntactic structures with time. The studies presented above provide strong indications that linguistic features in texts can afford important insights into writing quality and development. Importantly, it

seems that there is an increasing focus on linguistic features in writing and that new research is providing more robust and principled findings that can help guide the writing field.

Obviously, this narrative is a bit simplified. There are a number of intervening factors and research results that influence the generalized findings presented above that complicate linguistic analyses of texts. These complications can have important effects on writing analyses of which specialists and non-specialists should be aware. The complications arise from the interdisciplinary nature of this type of research which often combines writing, linguistics, statistics, and computer science fields. With so many fields involved, it is often easy to overlook confounding factors. Among these factors are how the linguistic features are calculated as well as the limitations of focusing only on linguistic features, differences between L1 and L2 writing populations, differences between writing tasks, topics, prompts, and disciplines, the effects of individual differences and demographics, and the use of human judgments as metrics for writing quality. Lastly, as mentioned earlier, linguistic features do not work alone, but rather in conjunction with one another, especially when predicting writing quality. While this list is not exclusive, it provides a strong starting point for discussing considerations into textual analyses and writing analyses in general.

5.1 Calculating Linguistic Features

One problem with early studies of textual features is that the features were either entirely or mostly coded by hand, which is error prone, subjective, and time consuming. In the late 80s, NLP tools came into existence, but they were not freely accessible nor user-friendly. With time, NLP tools seem to have become the method of choice for linguistic analyses of writing for a number of reasons. Chief among them is convenience, because NLP tools can efficiently analyze massive amounts of data by repeating simple computations objectively and literally, something that is time consuming and difficult for humans to accomplish. However, with that convenience comes a number of caveats. NLP tools are based on simple computer programs that rely on a sequence of instructions that tell the program how to complete a task. NLP tools require, at some level, knowledge of language and that knowledge is almost guaranteed to be impoverished when compared to human knowledge. At best, the linguistic features reported by NLP tools are proxies for actual language knowledge and while these proxies continuously improve (consider the difference between counting the number of letters in a word versus calculating actual word frequency in a representative corpus as proxies for lexical sophistication), they are still imperfect. Additionally, a great deal of specialized background is needed to understand the foundational knowledge presented by NLP tools. Without this knowledge, conclusions reached based on NLP analyses may be misleading. NLP tools are generally not used in isolation as well, because the numbers reported by the majority of NLP tools are meaningless in the absence of inferential statistics or machine learning algorithms.

Thus, researchers willing to use these tools need to rely on multiple, specialized domains.

Perhaps a bigger limitation to current NLP tools is not what the can measure, but more importantly what they cannot measure, especially in terms of student writing. While most NLP tools can measure the presence of linguistic features, they cannot measure whether the features are used accurately whether in terms of context or form. For instance, a writer may produce the word "wit," which is an infrequent word, but the word may be produced in the sentence "Eating wit friends is important." The "wit" in this example is an obvious misspelling, but to the NLP tool, the writer would appear to have produced an infrequent word. Beyond accuracy, NLP tools generally only measure simple structures like words, phrases, and sentences, but not complex structures important to writing such as claims, arguments, and evidence.

More importantly, NLP tools cannot measure pragmatic information such as argumentation, flow, or style. Often this means that it is difficult to know if the linguistic features measured relate to a text's conceptual content or writing style. For instance, it is difficult to assess whether the lexical diversity of the text is measuring writers' vocabulary knowledge, purposeful lexical repetition meant to increase text cohesion (i.e., content), or stylistic choices (i.e., the use of synonyms). Expectational differences in how some lexical features interact with writing quality also highlight inconsistencies in content and stylistic interpretations of linguistic features. As an example, it is generally assumed that more advanced writers will produce more sophisticated words. However, some studies indicate that more advanced writers produce less sophisticated lexical items. Specifically, McNamara et al. (2013) reported that higher rated essays included more specific words and more imageable words, while Crossley et al. (2011) found that essays written at a higher grade level contained more concrete words and words with fewer senses. Conceptually this make sense because advancing writers are likely providing more specific evidence to support claims, which may manifest itself lexically as more imageable, concrete, and specific words. Stylistically, however, it may be expected that writers would produce less imageable, concrete, and specific words in order to demonstrate mastery of the lexicon and write more "academically."

With time, these limitations may be addressed with advances in machine learning and computational linguistics, especially as NLP analyses become more common. A surge in user-friendly and freely accessible NLP tools within the last 15 years has allowed for a surge of textual analyses. Many of these tools have been developed for non-specialist in order to increase access to NLP analyses (e.g., Crossley, Kyle & Dascalu, in press; Graesser, McNamara, Louwerse, & Cai, 2004; Kyle & Crossley, 2017; Kyle, Crossley, & Berger, 2018). Most of these tools work only with the English language, but some tools are multilingual (MacWhinney, 2014, Dascalu, Trausan-Matu, McNamara, Dessus, 2015). Linguistically, the tools can provide information about text cohesion, lexical attributes of a text, syntactic complexity metrics, and emotion and

affective features, all of which can be used to better understand writing quality and development. However, researchers need to know their limitations and how to interpret the tools' output.

5.2 First Language and Second Language Writing

Linguistic analyses of writing appear to be more common in second language (L2) studies for two likely reasons, 1) the majority of L2 writing researchers are linguists while the majority of L1 writing researchers are not, and 2) analyses of L2 writing are often used to examine language proficiency and not necessarily writing proficiency. Thus, much of the research reported in this paper focuses on L2 writers and the evidence seems to indicate similar trends between L1 and L2 writing in terms of assessments of quality and development, especially in terms of lexical and syntactic features. In general, like L1 writing, higher quality L2 writing and more developed L2 writing contains greater lexical sophistication and syntactic complexity. Additionally, with advanced writers, there may be no differences between L1 and L2 writers in terms of overall writing quality (Attali & Powers, 2008),

However, L2 writers differ from L1 writers in a few important ways that merit discussion. First, L2 writers vary significantly in their language proficiency, whereas most L1 writers have similar proficiency levels (i.e., they are all fluent in their native language). Differences in language proficiency among L2 writers should be controlled for in any textual analyses. Second, many L2 writers are already literate when they begin to learn a new language and the literacy they have in their L1 can transfer to their L2, especially their knowledge of writing strategies. The transfer of writing strategies may influence linguistic features, especially those related to text cohesion, and pragmatic functions such as argument structure and style. Thus, comparison across L1 and L2 populations should be done with care as should analyses that include both L1 and L2 writers. Additionally, there is some evidence that native versus non-native speaking status may affect human ratings of writing quality differently. For instance, in terms of phrasal production, L1 writers that produce more complex phrases (i.e., less frequent phrases and a small proportion of common phrases) are scored higher in terms of writing quality (Crossley et al., 2012). This contrasts with L2 studies which show that L2 writing is judged to be of higher quality if it contains more frequent bi-grams and a greater proportion of common phrases (Kyle & Crossley, 2015; Garner et al., 2018). These differences may result from raters' perceptions of writers' native language status wherein for advanced L2 writers they favor texts that demonstrated adherence to expected norms because beginning L2 writers produce more infrequent and less common phrases that may also be ungrammatical as a result of lacking phrasal knowledge. In contrast, advanced L1 writers produce less frequent and common phrases while lower level L1 writers have the background knowledge to produce expected phrases.

Lastly, L2 writers may differ in terms of both linguistic and orthographic distances to the language they are learning. For instance, you would expect that Chinese learners of English would have a more difficult job writing in English than a German learner of English because Chinese lacks similarities in linguistic structures and vocabulary (because German and English belong to the same language families) as well as orthographies (German uses an alphabetic system with similar orthography to English while Chinese uses a character-based writing system). These factors may influence both writing quality and developmental patterns and writing studies should attempt to control for language and orthographic differences. One way to do this is include language background as a predictor in statistical analyses. However, too many language backgrounds will make interpretation difficult, so researchers may want to use a continuous variable for language distances as that proposed by Chiswick and Miller (2004), noting its limitations.

5.3 Writing Tasks

Another consideration when assess linguistic features in writing samples involves the writing task itself with the understanding that different writing tasks may require different linguistic skills (Plakans, 2008; Plakans & Gebril, 2013). For instance, early studies into differences between expository and narrative writing indicated that expository texts contained less lexical repetition (Berman & Verhoeven, 2002) and that syntactic coordination was more common in narratives while syntactic subordination was more common in expository texts (Verhoeven et al., 2002). Most early studies of writing focused on independent writing samples where writers were generally expected to produce a classic five paragraph essay within a specific timeframe (generally around 25 minutes). More recent studies have begun to focus on source-based writing, which is considered a more authentic writing task because it is often used within academic settings. A number of studies examining both independent and source-based writing samples produced by the same writers indicates significant differences in linguistic output between the tasks. For instance, Guo et al. (2013) found that lexical sophistication features were significant predictors of writing quality for both independent and source-based writing, but local cohesion features were only predictive of source-based writing. Kyle and Crossley (2016) found that lexical range and bigram features were predictive of independent writing quality but not source-based writing even though source-based writing included more sophisticated lexical items.

Another concern is the overall number of linguistic studies that have been conducted on independent writing samples as compared to those of other writing tasks including source-based writing, summarizations, and narratives (although there were a number of early studies on narratives; Fitzgerald & Spiegel, 1986; Yde & Spoelders, 1985; Zarnowski, 1983). While linguistic studies of textual features in these lesser studied writing tasks are becoming more common (Crossley et al., 2019; Jorge-Botana, Luzón, Gómez-Veiga, & Martín-Cordero, 2015; Li, Cai, & Graesser, 2018; Mintz,

Stefanescu, D'Mello, & Graesser, 2014; Somasundaran et al., 2018), the majority of information available to the field about the interactions between writing quality/development and linguistic features is derived from a single task (independent writing), shedding some confidence on the generalizability of the findings to other tasks. Other concerns regarding task involve potential differences between writing samples collected from standardized tests and writing samples from the classroom and differences between timed and untimed writing. It is likely that writing for standardized tests (e.g., SAT, GRE, or TOEFL samples) may produce different linguistic features than writing samples produced in the classroom or in other, more authentic writing environments (e.g., business proposals, e-mails, journaling, blogs, or research reports). Linguistic differences are also likely in timed versus untimed writing samples. Timed samples provide fewer opportunities for planning, revising, and idea development as compared to untimed samples where students are more likely to plan, reflect, and revise writing. These differences may surface in timed writing such that it would be less cohesive and less complex both lexically and syntactically.

5.4 Topic and Prompt Effects

Another important consideration in linguistic analyses of writing samples is in terms of topic and/or prompt. Multiple studies have demonstrated that differences in topic or prompt can lead to the production of different linguistic features (Crossley et al., 2011; Hinkel, 2002; Huot, 1990; Tedick, 1990). Much of this has to do with linguistic priming in that writers are likely to be primed by words in the prompt to either copy linguistic forms and structures or produce related forms and structures. In either case, some of the linguistic features produced by writers may not represent their knowledge, but rather wording in the prompt. As an example, a prompt that asks respondents to write about global warming would likely produce more sophisticated lexical items than a prompt that asks respondents to write about their favorite animal based solely on the type of lexical knowledge writers will be asked to produce (Hinkle, 2002). Respondents may also mimic the structures in the prompt such that a more syntactically complex prompt prime more complex structures in the response (Tedick, 1990; Hinkel, 2002). The same can occur for cohesion features (Crossley, Varner, & McNamara, 2013). It is very likely that some prompts will promote greater content based on the concepts they contain while others will influence writing style by the nature of their wording. With this in mind, it is increasingly important to control for prompt differences in NLP analyses.

5.5 Discipline Differences

While uncommon, a few studies have demonstrated that differences in linguistic structures may also be discipline based. For example, Durrant (2017) found differences in vocabulary use between soft sciences (e.g., law, English, classics) and hard sciences (engineering, chemistry, biological sciences). Ward (2007) found differences within a

single discipline such that collocation use differed among 5 different engineering disciplines. Lastly, Crossley, Russell, Kyle, & Römer (2017) reported linguistic differences in writing between macro-disciplines (science and engineering) and micro-disciplines (biology, physics, electrical engineering, and mechanical engineering) at the lexical, syntactic, and cohesion levels. Thus, even differences in disciplines needs to be controlled for in linguistic writing analyses lest reported differences in linguistic features be misinterpreted.

5.6 Individual Variation

Another consideration when assessing linguistic features and their interaction with writing quality and/or development is individual variation on the part of the writer. For instance, previous studies have demonstrated that writing quality is correlated with stronger reading skills (Allen, Snow, Jackson, Crossley, & McNamara, 2014; Fitzgerald & Shanahan, 2000; Tierney & Shanahan, 1991), greater vocabulary knowledge (Allen & McNamara, 2014; Allen, Snow, Crossley et al., 2014; Stæhr, 2008), grade level (Attali & Powers, 2008), greater flexibility (Allen, Snow, & McNamara, 2014; 2016), and more writing-specific knowledge (Saddler & Graham, 2007). Demographically, Attali & Powers (2008) reported that females scored higher than males with a small negative interaction with grade level and that Asian and White students scored better than non-White students. At least one study has combined individual variation and lexical features to predict writing quality (Crossley, Allen, Snow, & McNamara, 2016). This study found that both linguistic features and an individual variation measure (reading ability) led to gains in predicting essay scores. Again, studies considering interactions with linguistic features and writing quality or development need to consider individual differences within the sampled population.

5.7 Interactions among Linguistic Features

Most investigations into writing quality have not focused on single linguistic constructs alone (i.e., lexical sophistication, syntactic complexity, or textual cohesion). Instead, linguistic writing research examines multiple linguistic constructs and multiple features from each construct at the same time to predict writing quality. Thus, unlike the review of textual feature above, writing quality is not generally investigated using isolated linguistic constructs but rather examining how construct comprised of multiple linguistic features interact.

As an example, with L1 writers, McNamara et al (2010) used two lexical features and one syntactic feature to predict essay score and found that the three features explained around 22% of the variance. Specifically, they reported the majority of the variance was explained by a syntactic complexity index (~12%) while the two lexical variables explained around 10% of the variance. McNamara et al. (2015) examined differences in low and high quality essays based on length. They found that multiple linguistic features informed their final model, which predicted exact scoring matches

55% of the time and adjacent scoring matches 92% of the time. They reported that cohesion indices were the strongest predictors of short essays that were of low quality. For longer essays, lexical, syntactic, and cohesion indices were all important predictors of lower quality essays while higher quality essays were best predicted by lexical and cohesion features. Crossley, Kyle, & McNamara (2015) used component features (i.e., components developed using multiple indices) to assess essay quality and found that three components (text length, lexical sophistication, and global cohesion) explained 40% of the variance in writing quality. They also reported that lexical sophistication explained a greater amount of variance than global cohesion.

For L2 writers, Crossley and McNamara (2012) found that five variables related to lexical sophistication and cohesion predicted 26% of the variance in writing quality with the majority of variance explained by lexical variables. Guo et al. (2013) used text features to predict both independent and integrated writing quality in the Test of English as a Foreign Language (TOEFL) writing samples. For the integrated essays, Guo et al. reported that seven features including lexical sophistication and cohesion features explained 58% of the variance in essay scores with the strongest predictors being lexical sophistication features (as compared to cohesion features). For the independent essays, Guo et al. found five features including lexical sophistication and cohesion indices explained 65% of the variance with lexical variables explaining greater variance than cohesion variables.

There is also evidence of multiple different profiles of successful writing that may depend on different, interactional linguistic feature sets. For L1 writing, Crossley et al. (2014) examined the linguistic profiles of successful essays using cluster analyses and found four distinct profiles. Specifically, the linguistic features in higher quality essays were comprised of features related to four writing styles: *action and depiction style, academic style, accessible style*, and *lexical style*. However, all the writing styles could lead to successful essays. An earlier study conducted by Jarvis, Grant, Bikowski, and Ferris (2003) used a similar approach for L2 writers. In their study they found that successful L2 writing also consisted of multiple different linguistic profiles. For instance, they found differences in high quality essays in terms of word length, noun and pronouns use, the use of the present tense and adverbials, and syntactic complexity. Both of these studies indicate that linguistic can features interact with one another to produce higher quality essays and that successful writing cannot be defined by a fixed set of linguistic features.

6. Conclusion

This paper has provided an overview of how linguistic features in writing samples can be used to estimate and predict writing quality and development. The goal of the paper is to demonstrate the strengths and limitations of linguistic approaches to writing research and discuss the growth of linguistic analyses as a result of rising interest in NLP

tools. Importantly, the paper provides a guide of potential pitfalls in linguistic analyses of writing samples.

While there are a number of potential limitations to linguistic analyses of writing, advanced NLP tools and programs have begun to address linguistic complications while better data collection methods and more robust statistical and machine learning approaches can help to control for confounding variables such as first language differences, prompt effects, and variation at the individual level. This means that we are slowly gaining a better understanding of interactions between linguistic production and text quality and writing development across multiple types of writers, tasks, prompts, and disciplines. Newer studies are beginning to also look at interaction between linguistic features in text (product measures) and writing process characteristics such as fluency (bursts), revisions (deletions and insertions) or source use (Leijten & Van Waes, 2013; Ranalli, Feng, Sinharry, & Chukharev-Hudilainen, 2018; Sinharay, Zhang, & Deane, 2019). Future work on the computational side may address concerns related to the accuracy of NLP tools, the classification of important discourse structures such as claims and arguments, and eventually even predictions of argumentation strength, flow, and style.

Importantly, we need not wait for the future because linguistic text analyses have immediate applications in automatic essay scoring (AES) and automatic writing evaluation (AWE), both of which are becoming more common and can have profound effects on the teaching and learning of writing skills. Current issues for both AES and AWE involve both model reliability (Attali & Burstein, 2006; Deane, Williams, Weng, & Trapani, 2013; Perelman, 2014) and construct validity (Condon, 2013; Crusan, 2010; Deane et al., 2013; Elliot et al., 2013, Haswell, 2006; Perelman, 2012), but more principled analyses of linguistic feature, especially those that go beyond words and structures, are helping to alleviate those concern and should only improve over time. That being said, the analysis of linguistic features in writing can help us not only better understand writing quality and development but also improve the teaching and learning of writing skills and strategies.

Notes

1. I guide readers interested in recent developments in AES and AWE systems to Strobl et al. (2019), who wrote a thorough review of language technologies designed to support writing instruction in secondary and higher education.
2. This paper will not provide an overview of linguistic studies examining genre variation or variation between L1 and L2 writers (e.g., Chen & Baker, 2010; Hyland, 2008; Nesi & Gardner, 2012) because the focus of this paper is on relations between linguistic features and both writing quality and writing development.
3. Writing may not adhere to linear patterns (Purves, 1992) and that are likely multiple patterns that lead to writing success (Crossley, Roscoe, & McNamara, 2014).

Acknowledgements

I would like to state a tremendous debt to all my colleagues including, but not limited to, Danielle McNamara, Kristopher Kyle, Laura Allen, Art Graesser, Carl Cai, Minkyung Kim, and Stephen Skalicky. I am also thankful to Luuk Van Waes for pushing this paper forward and to the anonymous reviewers who provided feedback on early draft.

References

Allen, L. K., & McNamara, D. S. (2014). You are your words: Modeling students' vocabulary knowledge with natural language processing. Manuscript submitted to the *8th International Conference on Educational Data Mining (EDM 2015)*.

Allen, L. K., Snow, E. L., Jackson, G. T., Crossley, S. A., & McNamara, D. S. (2014). Reading components and their relation to writing. *L'Année psychologique/Topics in Cognitive Psychology*. *114* (4), 663-691. https://doi.org/10.4074/s0003503314004047

Allen, L. K., Snow, E. L., & McNamara, D. S. (2014). The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, S. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), Proceedings of the 7th International Conference on Educational Data Mining (pp. 304-307). London, UK. Allen, Snow, & McNamara, 2016

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater1 V. 2. The Journal of Technology, Learning, and Assessment 4(3) . http://ejournals.bc.edu/ojs/index.php/jtla/index.

Attali, Y., & Powers, D. (2008). A developmental writing scale. ETS Research Report Series, 2008(1). Princeton, NJ: ETS https://doi.org/10.1002/j.2333-8504.2008.tb02105.x

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. doi:10.3758/BF03193014

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly,* 26, 390-395. https://doi.org/10.2307/3587016

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). New York, NY: Guilford Press

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

Berman, R. and Verhoeven, L. (2002). Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. *Written Language and Literacy, 5* (1), 1-43. https://doi.org/10.1075/wll.5.1

Berninger, V., Fuller, F., & Whitaker, D. (1996). A process approach to writing development across the life span. *Educational Psychology Review, 8,* 193-218. *https://doi.org/10.1007/bf01464073*

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. https://doi.org/10.1007/s11145-010-9262-y

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

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45* (1), 5-35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, G. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education. https://doi.org/10.1017/s0022226702211627

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

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. Linguistics and Education, 20, 50–66. https://doi.org/10.1016/j.linged.2009.01.005

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. https://doi.org/10.1017/s0142716400007293

Casanave, C. (1994). Language development in students' journals. *Journal of Second Language Writing, 3*, 179–201. doi:10.1016/1060-3743(94)90016-7.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*, 30–49.

Chiswick, B. R., & Miller, P. W. (2005). Linguistic distance: A quantitative measure of the distance between English and other languages. *Journal of Multilingual and Multicultural Development*, *26*(1), 1-11. https://doi.org/10.1080/14790710508668395

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*, 100-108. https://doi.org/10.1016/j.asw.2012.11.001

Connor, U. (1990). Linguistic/rhetorical measures for international student persuasive writing. *Research in the Teaching of English, 2* (4), 67-87.

Cox, B.E., Shanahan, T. & Sulzby, E. (1991). Good and poor elementary reader's use of cohesion in writing. *Reading Research Quarterly*. 26, 47–65. https://doi.org/10.2307/747987

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34,* 213-238. https://doi.org/10.2307/3587951

Crossley, S. A., Allen, L., Snow, E., & McNamara, D. S. (2016). Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *Journal of Educational Data Mining, 8* (2), 1-19.

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., Kim, M., Allen, L., & McNamara, D. S. (2019). Modeling Text Summarization Skills Using Natural Language Processing Tools. *Proceedings of the 20th International Conference on Artificial Intelligence in Education*.

Crossley, S. A., Kyle, K., Varner, L., Gou, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment, 7* (1).

Crossley, S. A., Kyle, K., & Dascalu, M. (in press). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating Semantic Similarity and Text Overlap. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1142-4

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). Austin, TX: 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). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35* (2), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

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. https://doi.org/10.1016/j.jslw.2014.09.006

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. https://doi.org/10.17239/jowr-2016.07.3.02

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, and A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education*. (pp. 438-440). New York: Springer. https://doi.org/10.1007/978-3-642-21869-9_62

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication, 31* (2), 184-215. https://doi.org/10.1177/0741088314526354

Crossley, S. A., Russell, D., Kyle, K., & Römer, U. (2017). Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics, 1,* 48-81.

Crossley, S. A., & Skalicky, S. (in press). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley, and McNamara (2011). *Language Teaching*. https://doi.org/10.1017/s0261444817000362

Crossley, S. A., Varner, L., & McNamara, D. S. (2013). Cohesion-based prompt effects in argumentative writing. In McCarthy, P. M. & Youngblood G. M., (Eds.). *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. (pp. 202-207). Menlo Park, CA: The AAAI Press.

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).

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. https://doi.org/10.1177/0741088311410188

Crowhurst, M. (1987). Cohesion in argument and narration at three grade levels. *Research in the Teaching of English, 21* (2), 185–197.

Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor, MI: University of Michigan Press.

Dascalu, M., Trausan-Matu, S., McNamara, D. S., Dessus, P. (2015). ReaderBench – Automated evaluation of collaboration based on cohesion and dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395-423. https://doi.org/10.1007/s11412-015-9226-y

Deane, P., Williams, F., Weng, V. Z., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *Journal of Writing Assessment,* 6(1), 40-56.

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

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics. https://doi.org/10.1093/applin/amv011*

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics*, *47*, 157–177. https://doi.org/10.1515/iral.2009.007

Englert, C. S., & Hiebert, E. H. (1984). Children's developing awareness of text structures in expository materials. *Journal of Educational Psychology*, *76* (1), 65. https://doi.org/10.1037//0022-0663.76.1.65

Elliot, N., Gere, A. R., Gibson, G., Toth, C., Whithaus, C., & Presswood, A. (2013). Uses and limitations of automated writing evaluation software. *WPA-CompPile Research Bibliographies, 23*.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44. https://doi.org/10.1017/s0267190512000025

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155. https://doi.org/10.1016/1060-3743(95)90004-7

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). Rowley, MA: Newbury House.

Fellbaum, C. (1998). *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*, 414–420. doi:10.2307/3587446.

Fitzgerald, J., & Spiegel, D. L. (1986). Textual Cohesion and Coherence in Children's Writing. *Research in the Teaching of English, 20*, 263-80.

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist, 35*(1), 39–50. doi: http://dx.doi.org/10.1207/S15326985 EP3501_5

Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multidimensional analysis. *Journal of Second Language Writing*, *26*, 80–95. http://doi.org/10.1016/j.jslw.2014.09.007

Garner, J.R., Crossley, S.A., & Kyle, K. (2018). Beginning and intermediate L2 writer's use of n-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching. https://doi.org/10.1515/iral-2017-0089*

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

Graesser, A.C., McNamara, D.S., Louwerse, M.M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36, 193–202. https://doi.org/10.3758/bf03195564

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

Grant, L. & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9, 123–145. https://doi.org/10.1016/s1060-3743(00)00019-9

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

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

Haswell, R. H. (1986) *Change in Undergraduate and Post-Graduate Writing Performance: Quantified Findings*. ERIC: 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

Haswell, R. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17* (3), 307–352. https://doi.org/10.1177/0741088300017003001

Haswell, R. H. (2006). Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In: P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 57-78). Logan, UT: Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p.7

Hayes, J., & Flower, L. (1980) Identifying the organization of writing processes. In Gregg, Lee; Steinberg, Erwin (eds.) *Cognitive processes in writing: An interdisciplinary approach*. Hillsdale, NJ: Lawrence Erlbaum, 3–30. https://doi.org/10.1017/s0142716400006585

Hinkel, E. (2002). *Second language writers' text*. Mahwah, NJ: Lawrence Erlbaum Associates.

Huang, K. (2015). More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. *System*, *53*, 13–23. https://doi.org/10.1016/j.system.2015.06.011

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

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

Hunt, K. (1970). Syntactic maturity in schoolchildren and adults. *Monographs of the society for research in child development, 35* (1), 1–67. https://doi.org/10.2307/1165818

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237-263. https://doi.org/10.2307/1170611

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4 (1), 51-69. https://doi.org/10.1016/1060-3743(95)90023-3

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

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

Jin, W. (2001). *A quantitative study of cohesion in Chinese graduate students' writing: Variations across genres and proficiency levels*. Retrieved from ERIC database (ED452726).

Jorge-Botana, G., Luzón, J. M., Gómez-Veiga, I., & Martín-Cordero, J. I. (2015). Automated LSA assessment of summaries in distance education: Some variables to be considered. *Journal of Educational Composition Research, 52* (3), 341–364. https://doi.org/10.1177/ 0735633 115571930

Jung, J., Crossley, S. A., & McNamara, D. S. (in press). Predicting Second Language Writing Proficiency in Learner Texts Using Computational Tools. *The Journal of Asia TEFL*. https://doi.org/10.18823/asiatefl.2019.16.1.3.37

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. https://doi.org/10.1044/0161-1461.1603.187

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA. https://doi.org/10.1111/modl.12468

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly, 49* (4), 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*(4), 12-24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34* (4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *Modern Language Journal, 102* (2), 333-349. https://doi.org/10.1111/modl.12468

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

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

Larsen–Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*, 439–448. https://doi.org/10.2307/3586142

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322. https://doi.org/10.1093/applin/16.3.307

Laufer, B., & Waldman, T. (2011). Verb-Noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647-672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lenko-Szymanska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics*, *19*, 225–251. https://doi.org/10.1075/ijcl.19.2.04len

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Li, H., Cai, Z., & Graesser, A. C. (2018). Computerized summary scoring: Crowdsourcing-based latent semantic analysis. *Behavior Research Methods*, *50*(5), 2144–2161. https://doi.org/10.3758/s13428-017-0982-7

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*, 85–102. https://doi.org/10.1016/j.jslw.2009.02.001

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

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

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. https://doi.org/10.1007/s11145-018-9853-6

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press. https://doi.org/10.4324/9781315805641

Masters, G., & Forster, M. (1996). *Progress maps (Part of the Assessment Resource Kit)*. Melbourne, Australia: The Australian Council for Educational Research.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International, 66*(12), (UMI No. 3199485)

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

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25*, 431-444. https://doi.org/10.1016/0749-596x(86)90036-7

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

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*, 295-323. https://doi.org/10.1177/00238309010440030101

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. Written Communication, 27 (1), 57-86. https://doi.org/10.1177/0741088309351547

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. https://doi.org/10.3758/s13428-012-0258-1

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 https://doi.org/10.1207/s1532690xci1401_1

Mintz, L., Stefanescu, D., D'Mello, S. K. D., & Graesser, A. C. (2014). Automatic Assessment of Student Reading Comprehension from Short Summaries. *Proceedings of the 7th International Conference on Educational Data Mining*, 333–334.

Myhill, D.A. (2008). Towards a Linguistic Model of Sentence Development in Writing. *Language and Education*, *22* (5), 271-288. https://doi.org/10.1080/09500780802152655

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

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

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, E. A. Moravcsik, H. Ouali & K. M. Wheatley (Eds.), Formulaic Language (Volume 2): Acquisition, Loss, Psychological Reality, and Functional Explanations (pp. 375-385). Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.83.07ohl

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

Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. *Second Language Research*. https://doi.org/10.1177/0267658317694221.

Perera, K. (1984) Children's writing and reading: analysing classroom language (Oxford, Basil Blackwell).

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In: C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-131). Fort Collins, Colorado: WAC Clearinghouse/Anderson, SC: Parlor Press.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21,* 104-111. https://doi.org/10.1016/j.asw.2014.05.001

Perin, D., & Lauterbach, M. (2016). Assessing text-basedwriting of low-skilled college students. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-016-0122-z.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111–129. https://doi.org/10.1016/j.asw.2008.07.001

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217–230. https://doi.org/10.1016/j.jslw.2013.02.003

Popham, J. (2007). The lowdown on learning progressions. *Educational Leadership*, *64*, 83–84.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English, 26* (1), 108–122.

Ranalli, J., Feng, H.-H., & Chukharev-Hudilainen, E. (2018). Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing.*Assessing Writing, 36*, 77-89. https://doi.org/10.1016/j.asw.2018.03.007

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective*. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.

Saddler, B., & Graham, S. (2007). The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading and Writing Quarterly, 23,* 231-247. https://doi.org/10.1080/10573560701277575

Salsbury, T., Crossley, S. A, & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27* (3), 343-360. DOI: 10.1177/0267658310395851.

Sanders, T. J., & Pander Maat, H. (2006). Cohesion and coherence: Linguistic approaches. In Encyclopedia of Language & Linguistics (pp. 591-595). https://doi.org/10.1016/b0-08-044854-2/00497-1

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech. Studies in Second Language Acquisition, 38, 677–701. https://doi.org/10.1017/s0272263115000297

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

Sinharay, S., Zhang, M., & Deane, P. (2019) Prediction of Essay Scores From Writing Process and Product Features Using Data Mining Methods, *Applied Measurement in Education, 32*(2), 116-137. https://doi.org/10.1080/08957347.2019.1577245

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

Somasundaran, S., Flor, M., Chodorow, M., Molloy, H., Gyawali, B., & Mcculla, L. (2018). Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, *6*(1), 91–106. https://doi.org/10.1162/TACL_A_00007

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139–152. doi: 10.1080/09571730802389975

Stockwell, G., & Harrington, M. (2003). The incidental development of L2 proficiency in NS-NNS email interactions. CALICO Journal, 20, 337–359. https://doi.org/10.1558/cj.v20i2.337-359

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist, Assessing Writing, 18, 187-201. https://doi.org/10.1016/j.asw.2013.05.001

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. TESOL Quarterly, 47, 420–430. https://doi.org/10.1002/tesq.91

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. English for Specific Purposes, 9, 123–143. https://doi.org/10.1016/0889-4906(90)90003-u

Tierney, R. J., & Shanahan, T. (1996). Research on the relationship: Interaction, transactions, and outcomes. *Handbook of Reading Research*, *2*, 246.

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. https://doi.org/10.1075/wll.5.2.02ver

Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in Skills for Life Writing examinations. *Cambridge ESOL: Research Notes*, *41*, 7–14.

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. https://doi.org/10.1007/s11145-010-9266-7

Ward, J. (2007), "Collocation and Technicality in EAP Engineering," Journal of English for Academic Purposes, 6 (1): 18–35. https://doi.org/10.1016/j.jeap.2006.10.001

Wilson, M., & Bertenthal, M. (Eds.). (2005). *Systems for state science assessment. Board on Testing and Assessment, Center for Education, National Research Council of the National Academies*. Washington, DC: National Academies Press.

Witte, S., & Faigley, L. (1981). Coherence, cohesion, and writing quality. College Composition and Communication, 32, 189-204. https://doi.org/10.2307/356693

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). Second language development in writing: Measures of fluency, accuracy, and complexity. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center. https://doi.org/10.1017/s0272263101263050

Yde, P., & Spoelders, M. (1985). Text cohesion: An exploratory study with beginning writers. Applied Psycholinguistics, 6 (4), 407–415. https://doi.org/10.1017/s0142716400006330

Zarnowski, M. (1983). Cohesion in student narratives: Grades four, six, and eight. Unpublished research report (ERIC Document Reproduction Service No. ED 247 569).