Journal of Second Language Writing 32 (2016) 1–16

Contents lists available at ScienceDirect

Journal of Second Language Writing

The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality

Scott A. Crossley [a,*], Kristopher Kyle [a], Danielle S. McNamara [b]

[a] Georgia State University, Atlanta, GA 30303, USA
[b] Arizona State University, Tempe, AZ 85287, USA

ARTICLE INFO

Article history:
Received 2 November 2014
Received in revised form 26 January 2016
Accepted 27 January 2016

Keywords:
Cohesion
Second language acquisition
Second language writing
Essay quality
Natural language processing
Learner corpus

ABSTRACT

An important topic in writing research has been the use of cohesive features. Much of this research has focused on local and text cohesion. The few studies that have studied global cohesion have been restricted to first language writing. This study investigates the development of local, global, and text cohesion in the writing of 57 s language (L2) university students and examines the effects of these cohesion types on judgments of L2 writing quality Growth is observed in the use of a number of local, global, and text cohesive features across a semester-long upper-level English for Academic Purposes (EAP) course. Local, global, and text features also predicted whether an essay was written at the beginning or the end of the semester with an accuracy of 71%. In addition, the use of local, global, and text cohesive features explains 36% of the variance in human judgments of text cohesion and 42% of the variance in overall judgments of writing quality. This study has important implications for second language acquisition, writing development, and writing pedagogy.

1. Introduction

The use and effects of cohesive devices in student writing has been of interest for some time (McCutchen & Perfetti, 1982; Witte & Faigley, 1981), but their impact on essay quality is unclear. For instance, the presence of local cohesive devices (i.e., devices related to sentence level cohesion such as connectives or word overlap between sentences) in writing produced by adult first language (L1) writers is often associated with judgments of lower writing quality (Crossley & McNamara, 2010, 2011; Evola, Mamer, & Lentz, 1980; McCulley, 1985). In contrast to L1 writing studies, a number of studies examining adult second language (L2) writing report positive correlations between the presence of local cohesive devices and writing quality (Jafarpur, 1991; Yang & Sun, 2012). There are several unexplored explanations for these differential findings.

One such explanation rests on differences in links between writing quality and the production of local cohesive devices, global cohesive devices (i.e., devices related to cohesion between larger chunks of texts such as word overlap between paragraphs in a text), and text cohesive devices (i.e., devices related to cohesion across an entire text such as the ratio of pronouns to nouns [givenness] and word repetition [lexical diversity] in the text). Recent computational studies have reported differences between local and global cohesive devices and their relation to writing quality for L1 writers, with local cohesion negatively related to writing quality and global cohesion positively related to writing quality (Crossley & McNamara, 2011; Crossley, Roscoe, McNamara, & Graesser, 2011). No studies, to our knowledge, however, have explicitly

* Corresponding author. Tel.: +1 404 413 5179.
E-mail addresses: scrossley@gsu.edu (S.A. Crossley), kkyle3@student.gsu.edu (K. Kyle), dsmcnamara1@gmail.com (D.S. McNamara).

http://dx.doi.org/10.1016/j.jslw.2016.01.003
1060-3743/

examined differences between local, global, and text cohesive devices in L2 writing. Understanding differences between these types of cohesive devices in L2 writing may help to better explain L2 writing proficiency and differing expectations for L2 writers on the part of expert raters.

Beyond examining the relations between cohesive devices and writing quality, there has also been an interest in investigating the longitudinal development of cohesive devices for both L1 learners (Bereiter & Scardamalia, 1987; Berninger, Fuller, & Whitaker, 1996; Hayes & Flower, 1980; Myhill, 2008) and L2 learners (Crossley, Salsbury, & McNamara, 2010a; Crossley, Salsbury, McNamara, & Jarvis, 2010; Yang & Sun, 2012). However, more research concerning the development of cohesive devices has been conducted for L1 writers than L2 writers resulting in a paucity of available information about cohesion development in L2 learners. To our knowledge, studies examining the development of local, global, and text cohesive devices in L2 learners are infrequent, and none of these links the development of these cohesive devices with judgments of writing quality.

This study addresses these gaps by examining the development of local, global, and text cohesive devices in L2 learners in conjunction with examining the relations such developments have on human judgments of writing quality (both judgments of overall writing proficiency and more fine-grained judgments of text coherence). Such an approach affords the opportunity to examine not only growth in the use of cohesive devices by L2 learners, but also links between such growth and expert judgments of essay quality. To do so, we use computational indices of local, global, and text cohesive devices to examine how the production of cohesive devices change over time in L2 writers (i.e., longitudinal growth[1]) and how the use of cohesive devices are related to human ratings of L2 writing. The use of computational tools affords us the opportunity to investigate large corpora of texts for a greater number of cohesion indices, something that was not possible in past research.

1.1. Cohesion and coherence

An important distinction in cohesion studies is the difference between cohesion and coherence. Cohesion generally refers to the presence or absence of linguistic cues in the text that allow the reader to make connections between the ideas in the text. Generally these cues are local in nature, but they can also be based on global or text cohesion. Examples of local cohesion cues include overlapping words and concepts between sentences and explicit connectives such as *because*, *therefore*, and *consequently* (Halliday & Hasan, 1976). Examples of global cohesion cues include semantic and lexical overlap between paragraphs in a text (Foltz, 2007) such that words or ideas in one paragraph are repeated in subsequent paragraphs. In addition, cohesion can be measured at the text level (i.e., throughout an entire text). One example of this is *givenness* in which cohesion is measured across the text based on the number of words that are new (e.g., an initial noun referent) or given (noun referents that can be referred to pronominally). In general, global and text cohesion cues are more implicit than local cohesion cues. In contrast to cohesion, coherence refers to the understanding that the reader derives from the text (i.e., the coherence of the text in the mind of the reader). This coherence depends on a number of factors including cohesion cues and nonlinguistic factors such as prior knowledge and reading skill (McNamara, Kintsch, Songer, & Kintsch, 1996; O'Reilly & McNamara, 2007).

A number of studies have shown that cohesive devices are important indicators of text comprehensibility such that an increase in text cohesion generally leads to greater comprehension of a text (Crossley, Yang, & McNamara, 2014b; Gernsbacher, 1990; Crossley & McNamara, 2011). However, the facilitative effects for cohesive device are stronger for low-knowledge readers than high-knowledge readers (McNamara et al., 1996). In terms of the relation between cohesive devices and human judgments of coherence, the results are more nuanced. At least three studies have indicated that local and text cohesion are either not related or negatively related to human ratings of text coherence in both L1 and L2 writing (Bestgen, Lories, & Thewissen, 2010; Crossley & McNamara, 2010; Crossley & McNamara, 2011). In contrast, Crossley and McNamara (2011) reported that markers of global cohesion in L1 writing were positively related to expert judgments of text coherence. This finding is supported by L1 longitudinal studies that indicate that developing writers show advancements in their use of global cohesion by developing greater links between paragraphs (Bereiter & Scardamalia, 1987; Hayes & Flower, 1980).

1.2. Development in the use of cohesive devices: L1 and L2 learners

A number of studies have investigated the development of cohesive devices in L1 writers, but fewer have focused on L2 learners. For L1 writers, most studies have supported the notion that the use of cohesive devices increases as writers develop, especially in elementary and middle school. In general, as L1 writers develop, there is an increase in the use of cohesive devices to manipulate text level structures (Bereiter & Scardamalia, 1987). However, there are strong grade level effects for cohesion indicating that students at various levels use cohesive devices differently (Crowhurst, 1987; Fitzgerald & Spiegel, 1986; Yde & Spoelders, 1985). For instance, studies have shown that as early as the second grade, writers begin developing more cohesive writing through the use of local cohesion devices such as referential pronouns and connectives (King & Rentel, 1979). In addition, Rentel, King, Pettegrew, and Pappas (1983) reported an increase in lexical repetition across grades 1–4. These studies along with others demonstrate that for young writers, the distance between the ties used to create cohesion

[1] The notion of growth should be considered relative because, in many cases, growth is actually related to decreasing linguistic features. For instance, a decrease in frequent words over time would still indicate positive lexical development.

decreases over time such that referents are closer to one another, leading to more cohesive text (Fitzgerald & Spiegel, 1986; McCutchen & Perfetti, 1982; Yde & Spoelders, 1985). The increased use of such local cohesion cues continues until around the 8th grade (McCutchen & Perfetti, 1982) such that 8th grade essays contain more connectives and overlap between words and concepts across sentences than 6th grade essays (McCutchen, 1986). At a later stage, generally for older children who are more advanced writers, there is a movement away from the use of local cohesive devices and a movement toward the development of more complex syntactic constructions, which can be used to implicitly connect ideas (i.e., through modifications and embeddings: Crossley et al., 2011; Haswell, 2000; McCutchen & Perfetti, 1982). For example, Crowhurst (1987) reported a decrease in the use of temporal and causal conjunctions in narratives across 6th, 10th, and 12th graders. Overall, Crowhurst found that older students tend to use a greater variety, but not necessarily a greater number, of conjunctions than younger students. Crossley et al. (2011) also found that college freshmen wrote essays that contained less explicit cohesion (e.g., fewer positive logical connectives and less content word overlap between adjacent sentences) than 11th graders, who produced essays with less cohesion than 9th graders, indicating that at more advanced levels, the reliance on local cohesive devices is less frequent. Haswell (1986) also reported that graduate students used fewer cohesive and identical ties than undergraduate writers and Haswell (1990) found differences between freshman and junior writers and undergraduate and graduate writers for local specification of given information (i.e., cohesive ties that connect given and new information) and local coherence (logical connectives).

Studies that examine the use of cohesive devices over time for L2 learners are less frequent and those available focus on the development of cohesive devices in speech, which may not overlap with developments in writing. For instance, Crossley, Salsbury, and McNamara (2010b) investigated the longitudinal development of six L2 speakers over the course of the year and found increased semantic similarity among utterances as learners studied a second language (i.e., participants showed increased repetition of semantically related words across utterances as a function of time). Crossley, Salsbury, and McNamara (2010c) found that over the course of a year of study, the number of misunderstandings between L1 speakers and L2 learners decreased and that this decrease was associated with an increase in global and text cohesive devices related to causality (i.e., causal verbs and particles) and semantic similarity between utterances. Thus, as L2 learners' speech became more coherent to the L1 interlocutors, the L2 learners employed a greater amount of global and text cohesion in their speech. Lastly, while not a longitudinal study, Yang and Sun (2012) compared differences between the argumentative writing of second and fourth-year undergraduate Chinese learners of English as an L2. They reported differences in the number of local cohesive devices (pronouns, conjunctions, ellipsis, and lexical overlap) used based on proficiency level such that more advanced learners used a greater number of cohesive devices and used them more accurately.

Overall, research indicates that L1 learners begin to depend less on local cohesion cues in their writing as they develop with age and grade level. The few studies examining L2 learners indicate that local, global, and text cohesion features may increase as a function of proficiency although available longitudinal research is limited to spoken data. A better understanding of these differences could help researchers develop more robust theories of second language acquisition and L2 writing and assist writing instructors in constructing more accurate expectations of L2 learner growth in the classroom.

1.3. Cohesion features and human judgments of writing quality

A number of studies examining both L1 and L2 writers have been conducted to investigate the relations between cohesive devices and human judgments of writing quality. In terms of adult L1 writers[2], with the exception of one study (Witte & Faigley, 1981), studies generally report that local cohesive devices are either not positive indicators of essay quality or are negative predictors, while global cohesive devices demonstrate positive relations with essay quality. The findings for L2 writing is mixed. Some studies report positive relations between the use of cohesive devices and essay quality, while other studies report either no relations or negative relations with essay quality.

In an early study of L1 writers, Evola et al. (1980) found that local cohesive devices were not strong indicators of essay quality. Witte and Faigley (1981) compared the use of local cohesive devices in low and high level L1 writers (college freshmen). They reported that high-level writers used a greater number of certain types of cohesive devices (reference and conjunctions) than low-level writers. McCulley (1985) later found that the overall number of cohesive devices in a writing sample was not a good predictor of essay quality for L1 writers, although the production of specific cohesive devices (synonyms and hypernyms) was related to essay quality (though these features are also strongly linked to lexical sophistication). In a similar study, Neuner (1987) found that the overall number of cohesive devices produced by L1 writers did not distinguish low- and high-level essays, although cohesive chains across paragraphs did (i.e., global cohesion).

More recent studies using computational tools to measure the use of cohesive devices in L1 adult writers have demonstrated that local cohesion cues are either unrelated or negatively related to essay quality, while global cohesion cues can be positively related to essay quality. For example, McNamara, Crossley, & McCarthy (2010a), found no difference in essay quality as a function of local and text cohesive devices (i.e., word and semantic overlap, positive logical connectives, logical operators, negative temporal connectives) in a corpus of freshman college writings. Additional studies by Crossley

[2] Cohesive devices are a strong indicator of writing quality for L1 children, with many studies reporting links between the quality of a writing sample and the production of cohesive devices (Bereiter & Scardamalia, 1987; Cameron et al., 1995; Cox, Shanahan, & Sulzby, 1990; Englert & Hiebert, 1984; Hayes & Flower, 1980; Myhill, 2008; Struthers, Lapadat, & MacMillan, 2013).

and McNamara (2010, 2011) indicated that local cohesive devices such as semantic coreference, causal cohesion, spatial cohesion (i.e., cohesion related to the use of location nouns and motion prepositions, such as *house* and *into*), connectives and logical operators, anaphoric resolution, and word overlap along with text cohesive measures such as temporal cohesion (i.e., the repetition of tense and aspect) either did not correlate or correlated negatively with human ratings of text quality, even though human ratings of coherence best explained overall judgments of essay quality. However, global cohesive devices that measured semantic similarity between paragraphs demonstrated positive relations with essay quality (Crossley & McNamara, 2011; McNamara, Crossley, & Roscoe, 2013; Crossley et al., 2011).

For L2 writers, a number of studies show positive relations between essay quality and the production of local and text cohesive devices. For instance, Jafarpur (1991) found that the quality of essays written in English by undergraduate Iranian students was correlated with the number of cohesive ties and cohesive types used in the essays. Chiang (2003) explored essay quality using expert holistic scores and analytic ratings of essay quality. Chiang found that human ratings of essay quality were best explained by analytic ratings of cohesion and coherence taken from L1 and L2 expert raters. The strongest analytic rating for explaining overall essay quality was local in nature: transitions between sentences in the absence of conjunctions. Liu and Braine (2005) examined a number of local cohesive devices and reported that essay quality scores for undergraduate Chinese L2 writers moderately correlated with the total number of cohesive devices in the text (i.e., text cohesion). In addition, the number of lexical cohesive devices strongly correlated with essay quality. Yang and Sun (2012) also reported strong correlations between the total number of correctly used cohesive devices and essay quality for argumentative essays written by Chinese writers of English.

In contrast to these studies, recent studies using computational tools have indicated that local cohesion is negatively related to essay quality. For instance, in a study that examined the writing quality of independent essays (i.e., impromptu writing) produced by Hong Kong high school students, Crossley and McNamara (2012) found that local and text cohesive devices such as content word overlap between adjacent sentences, positive logical connectives, aspect repetition, and semantic similarity between sentences were negatively correlated with expert ratings of essay quality for Hong Kong high school students. One of the text cohesion devices (aspect repetition) was a negative predictor in a regression model that predicted 26% of the variance in the essay scores. In a similar analysis, Guo, Crossley, and McNamara (2013) used local cohesion indices to examine independent essay scores and source-based essay scores (i.e., writing that requires the use of reading and/or listening materials as stimuli for composing an essay) found in the Test of English as a Foreign Language (TOEFL). Guo et al. (2013) reported that indices of local cohesion (e.g., content word overlap, and conditional connectives) and text cohesion (e.g., aspect repetition) were negatively correlated with judgments of essay quality for the independent essays. However, for the source-based essays that heavily relied on text integration (i.e., from outside sources), local cohesive indices (e.g., semantic similarity between sentences and noun overlap between sentences) were positively correlated with essay quality and included in regression models that predicted essay scores.

Overall, these studies (with the exception of Witte & Faigley, 1981) indicate that expert ratings of L1 essay quality are either not predicted by local cohesive devices or negatively predicted by local cohesive devices. In contrast, there is some evidence that expert ratings of L1 writing quality may be predicted by global cohesive indices. For L2 writing, the results are mixed. Earlier studies report positive relations between cohesion features and essay quality and more recent studies using computational approaches report that more cohesive devices equate to lower scores of essay quality. Notably, the majority of the cohesive devices investigated in these L2 studies have been local in nature. Thus, the relations between global cohesive devices and essay quality have yet to be explored. Exploring these relations could facilitate a better understanding of how linguistic elements of writing interact with judgments of proficiency providing links to pedagogical interventions and writing strategy development.

2. Method

The purpose of this study is to assess cohesion development in L2 writers and how this development may be related to expert judgments of writing quality. To this end, we use a number of automated cohesion indices that measure both local and global cohesion to assess growth in descriptive essays written by L2 learners at the beginning, middle, and end of a semester long writing course. We complement this analysis by assessing the degree to which these cohesion indices predict the variance in human ratings of essay quality for essays written throughout the course. In doing so, we address two key questions: (1) Do L2 writers demonstrate development in their use of cohesive devices over the course of a semester (i.e., longitudinal growth), and, if so, (2) Are there similarities between the cohesive devices that show growth over the course of a semester and the cohesive devices in L2 essays that predict human ratings of writing proficiency?

2.1. Corpus

The corpus used in the analysis consists of 30-min descriptive essays written by university level students ($N = 57$) enrolled in upper-level English for Academic Purposes (EAP) courses at Michigan State University (levels 3 and 4 of an intensive English program). There were a total of four sets of two essay topics. The students chose from one of two topics at each time of writing and the topic sets were counterbalanced so that each topic set was equally sampled at the beginning, middle, and end of the semester data collections. See Appendix A for the topic sets. Each student wrote three essays, one at the beginning of the semester, one at the semester midpoint, and one at the end of the course. Each essay was written in a

six- to eight-week interval (i.e., there were six to eight weeks between each submission but students did not have six to eight weeks to write each essay). The final corpus included 171 essays (57 writers × 3 essays each). The essays averaged 335.351 (SD = 97.537) words and 5.388 (SD = 4.165) paragraphs in length. A number of previous studies have used this corpus to examine the development and effects of phraseological competence (Bestgen & Granger, 2014), syntactic complexity (Bulte & Housen, 2014), and linguistic variation (Friginal & Weigle, 2014) in L2 learners. For additional details regarding the collection and makeup of the dataset, we refer the reader to Connor-Linton and Polio (2014).

2.2. Human ratings

Two expert raters assessed the quality of each essay using a composition grading scale that required the raters to rate each essay on five different analytical features: content, organization, vocabulary, language use, and mechanics (see Connor-Linton & Polio 2014, for additional information about the grading scale and raters). The essay scoring was blind. The raters did not know the time frame in which the essays were written or that multiple essays were written by a single writer. After scoring, the analytic ratings were combined into an overall rating for each essay. Of interest for this study is the combined rating for each essay (i.e., the overall quality of the essay) and the organization rating, which includes assessments of cohesion. Briefly, the organization rating presumes that traits related to higher writing proficiency include excellent overall organization, excellent use of transition words, excellent connections between paragraphs, unity within paragraphs, a clear thesis and substantive introduction and conclusion. The first four properties are strongly related to text cohesion while the latter two are linked to text cohesion, but not exclusively (i.e., they also have links to presenting and supporting main ideas and claims in the paper). Interrater reliability between the two raters for the essays written by the 57 participants in this study was acceptable (see Table 1). Strong multicollinearity was reported between all ratings (see Table 2).

2.3. Selected cohesion indices

To report on text cohesion, we selected cohesion indices from Coh-Metrix and the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, in press), both advanced computational tools that measure cohesion (Crossley & McNamara, 2014; Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Graesser, McCarthy, & Cai, 2014). TAACO, in contrast to Coh-Metrix, provides a greater breadth of global cohesion indices. TAACO also provides synonym overlap indices and part of speech (POS) tagged cohesion indices. Coh-Metrix, unlike TAACO, uses Latent Semantic Analysis (LSA) to provide semantic overlap. While both tools report some measures that are redundant with one another, we only selected indices from each tool that measured unique cohesion features. We used an automatic approach to assessing text cohesion because it affords speed, flexibility, and reliability (Higgins, Xi, Zechner, & Williamson, 2011). An overview of all selected cohesion indices is presented in Table 3.

2.3.1. Local cohesion indices

We selected a number of local cohesion indices from Coh-Metrix and TAACO. These indices included measurements of sentence cohesion (overlap of syntactic tags and phrases), incidence of theoretical and rhetorical connectives, semantic similarity between sentences, synonym overlap between sentences, and lexical overlap between sentences. These are discussed below.

2.3.1.1. Connectives. Coh-Metrix provides indices on five categories of theoretically-based connectives (Halliday & Hasan, 1976; Louwerse, 2001): causal (*because, so*), contrastive (*although, whereas*), additive (*moreover, and*), logical (*or, and*), and temporal (*first, until*). Finally, Coh-Metrix contrasts positive (*also, moreover*) versus negative (*however, but*) connectives. Coh-Metrix also calculates the incidence of conjuncts, subordinating conjunctions in a text, *and*'s, and all connectives (i.e., a list of all connectives found in English). TAACO reports a number of connective indices that are based on rhetorical features (i.e., not based on theoretical perspectives but rather on rhetorical uses). These connectives include coordinating connectives, semi-coordinators, basic coordinators, complex subordinators, contrasts, and opposition (e.g., *on the contrary*). Since connectives are used to connect clauses, phrases, and/or sentences, the indices are considered local in nature.

Table 1
Interrater reliability for human ratings.

Feature	*r*
Content	0.82
Organization	0.70
Vocabulary	0.76
Language use	0.77
Mechanics	0.85
Overall score	0.88

Table 2
Correlations between rating scales (including combined overall score).

Rating	Organization	Vocabulary	Language use	Mechanics	Overall score
Content	0.86**	0.82**	0.77**	0.63**	0.90**
Organization		0.76**	0.74**	0.67**	0.90**
Vocabulary			0.84**	0.70**	0.91**
Language use				0.76**	0.91**
Mechanics					0.86**

** $p < 0.001$.

2.3.1.2. Lexical overlap (between sentences). Coh-Metrix considers four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap (McNamara, Louwerse, McCarthy, & Graesser, 2010b). Argument overlap measures how often two sentences share nouns with common stems while stem overlap refers to how often a noun in one sentence shares a common stem with the other word types in a second sentence. Finally, content word overlap calculates the number of shared content words (nouns, verbs, adjectives, and adverbs) between adjacent sentences. TAACO also counts a number of lexical overlap indices. These indices compute lemma (e.g., *wrote* and *writes* both belong to the lemma *write*) overlap between two adjacent sentences and between three adjacent sentences for all words, content words, and function words. TAACO also calculates average overlap scores that are sensitive to a word/lemma's POS (e.g., noun, verb, adjective, adverb, and pronoun). In addition to providing the number of lemmas that overlap between adjacent sentences, TAACO calculates whether there is any overlap between adjacent sentences (binary overlap) for these features.

2.3.1.3. Semantic similarity (between sentences). Coh-Metrix uses Latent Semantic Analysis (LSA) to measure the semantic co-referentiality of a text. LSA is a statistical representation of deeper world knowledge used to assess the level of semantic similarity between text segments (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). Specifically, Coh-Metrix measures similarity between adjacent sentences as well as similarities between verbs in adjoining sentences.

2.3.1.4. Synonym overlap (between sentences). TAACO calculates overlap between words and word synsets between sentences for nouns and verbs using the WordNet database. Unlike strict overlap indices, these indices measure overlap between semantically related words (e.g., the synset for *jump* contains the related words *leap, bound*, and *spring* among others).

2.3.1.5. Syntactic cohesion. Coh-Metrix measures syntactic similarity (i.e., syntactic cohesion) by examining the uniformity and consistency of the part of speech (POS), phrasal, and clausal constructions between sentences in the text. These indices are calculated by comparing the similarity between adjacent sentences for these constructions.

Table 3
Cohesion features, categorization, tool, and examples.

Feature	Cohesion type	Tool	Description	Example of high cohesion
Causal cohesion	Text	Coh-Metrix	Use of causal verbs and particles	He *kicks* the ball *because* he is angry
Connectives	Local	Coh-Metrix/TAACO	A number of theoretical and rhetorical lists of connectives	*First*, she was rich *and* happy
Givenness	Text	TAACO	Ratio of pronouns to nouns; incidence of demonstratives	The *man* was happy *he* had *that*
Lexical diversity	Text	Coh-Metrix/TAACO	Word repetition across a text	*The big* dog saw *the big* cat
Lexical overlap	Both local and global	Coh-Metrix/TAACO	Overlap between nouns, arguments, stems, content and function words, and POS tags (for both sentences and paragraphs)	The *sun* was bright. The day was *sunny*
Semantic similarity	Local, global, and text	Coh-Metrix	LSA cosine values to measure similarity between text segments (for both sentences, paragraphs, and across the text)	The *dog* was tired. So was the *cat*
Spatial cohesion	Text	Coh-Metrix	Use of motion and temporal verbs and prepositions	He *placed* the book *on* the table
Synonymy overlap	Both local and global	TAACO	Overlap of synonyms across sentences and paragraphs	The *animal* was small. It was a *cat*
Syntactic cohesion	Text	Coh-Metrix	Overlap of POS tags and phrasal categories between sentences	She throws the ball. He hits the ball
Temporal cohesion	Text	Coh-Metrix	Repetition of tense and aspect	He *went* to the store because she *asked* him

2.3.2. Global cohesion indices

We selected a number of global cohesion indices from Coh-Metrix and TAACO. These indices all measure cohesion across paragraphs and include LSA based indices, lexical overlap, and synonym overlap. These are discussed below.

2.3.2.1. Semantic similarity (between paragraphs). Coh-Metrix also uses LSA to measure the semantic co-referentiality between paragraphs in a text. Specifically, Coh-Metrix measures similarity between adjacent paragraphs and similarity between paragraph types (i.e., initial, middle, and final paragraphs).

2.3.2.2. Lexical overlap (between paragraphs). TAACO calculates a number of paragraph overlap indices. These indices compute lemma overlap between two adjacent paragraphs and between three adjacent paragraphs using the same features as the sentence overlap indices (e.g., average and binary lemma overlap, content word lemma overlap, function word lemma overlap, and POS-sensitive lemma overlap).

2.3.2.3. Synonym overlap (between paragraphs). TAACO calculates overlap between words and word synsets between paragraphs for nouns and verbs using the WordNet database.

2.3.3. Text cohesion indices

We selected a number of text cohesion indices from Coh-Metrix and TAACO. These indices measure cohesion at the overall text level without regard to sentence or paragraph distinctions (i.e., these indices compute average scores across a text). These indices include causal cohesion, temporal cohesion, spatial cohesion (all related to situation model development; Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998), givenness (incidence of demonstratives and pronoun to noun ratios), lexical diversity, and semantic similarity. These are discussed below.

2.3.3.1. Causal cohesion. Coh-Metrix calculates the level of causal cohesion in a text by measuring the ratio of causal verbs to causal particles (Graesser et al., 2004). The measure of causal verbs is based on the frequency count of main causal verbs identified through WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). The causal particles are counted based on a defined set of particles, such as *because* and *as a result*.

2.3.3.2. Spatial cohesion. Coh-Metrix assesses spatial cohesion with two forms of information: location information (verbs and prepositions) and motion information (verbs and prepositions; Dufty, Graesser, Lightman, Crossley, & McNamara, 2006). Both motion verbs and location nouns are identified through WordNet (Fellbaum, 1998; Miller et al., 1990).

2.3.3.3. Temporal cohesion. Temporal cohesion is measured by Coh-Metrix in three ways: aspect repetition, tense repetition, and the combination of aspect and tense repetition. Time is represented in text through two dimensions: tense (past, present, future) and aspect (in progress versus completed). With the use of these dimensions, Coh-Metrix calculates the consistency of tense and aspect across an entire text.

2.3.3.4. Givenness. A greater number of pronouns and demonstratives are used when information is given (McNamara et al., 2014). Thus, TAACO calculates the ratio of nouns to pronouns with the presumption that a greater ratio of pronouns will relate to more given information. TAACO also counts the incidence of demonstratives (i.e., *this, those, that*, and *these*), which are generally used with given information.

2.3.3.5. Lexical diversity. Measures of lexical diversity relate to the repetition of words by a writer. Coh-Metrix reports on a number of sophisticated LD indices that control for text length, including *MTLD* (McCarthy, 2005; McCarthy & Jarvis, 2010) and *D* (Malvern, Richards, Chipere, & Durán, 2004). TAACO calculates a number of different type-token ratio (TTR) indices beyond those found in Coh-Metrix. These include TTR based on all of the words in a text, only content words (e.g., nouns, verbs, adjectives, and adverbs), and only function words (e.g., pronouns, preposition, and determiners). Other TTR indices reported by TAACO are based on lemma counts for all words, content words, and function words. In addition, TAACO calculates TTR for bigrams (2-word sequences) and for trigrams (3-word sequences).

2.3.3.6. Semantic similarity (across text). Coh-Metrix uses LSA to measure semantic similarity between all sentences and paragraphs (i.e., family resemblance) across the entire text.

2.4. Statistical analysis

There is little previous research that suggests the strength of one cohesion index over another in either predicting language growth over time or predicting human ratings of essay quality. For this reason, we take a data-driven approach to our statistical analysis in which we rely on effect sizes for the variables to assign importance and order of entry within our statistical models. Our first research question is whether and how L2 writers' use of cohesive devices changes during a semester-long writing course. For this analysis, we removed any variables that were not normally distributed and then conducted within-subjects Analyses of Variance (ANOVA) using the remaining selected Coh-Metrix and TAACO indices

focusing on the initial, middle, and final essays written over the course of the semester ($N = 171$). The ANOVAs provided us with information about which syntactic indices demonstrated significant linear growth patterns. Those indices that demonstrated significant linear growth patterns were then entered into a Support Vector Machine (SVM) supervised learning algorithm to assess how well the indices predicted if an essay was written at the beginning of the semester or at the end of the semester. An SVM learning algorithm produces a statistical learning model using classification and regression analysis that assigns a probability to each text given a number of instances (i.e., what is the probability that a text was written at the beginning or the end of the semester based on the cohesive features found within that text; Witten, Frank, & Hall, 2011).

For the SVM learning model, we tested the predictive strength of the indices using a ten-fold-cross-validation (10-CV) analysis (Witten et al., 2011). In this analysis, we chose 10 fixed folds wherein a tenth of the observations in turn were left out for testing and the remaining 90% of the instances were used as the training set. We assess the accuracy of the model by testing its ability to predict the classification (the human rating) of the omitted instances. Such an approach affords us the opportunity to test the models generated by the SVM on an independent data set (i.e., on essays that are not used to train the model). If the 10-CV results demonstrate significant classification results as reported by a Chi-squared test, our level of confidence in the model increases supporting the extension of the analysis to external data sets.

Our second research question addressed whether cohesive devices were predictive of human ratings of essay quality. To answer this question, we conducted regression analyses to examine if the selected Coh-Metrix and TAACO indices of cohesion were predictive of human ratings of essay quality for both the *organization* analytic ratings and *combined* score ratings. For this analysis we used all of the rated essays in the analysis ($N = 171$). After removing any variables that were not normally distributed, we conducted Pearson Product Moment Correlations between the human ratings for the essay and the cohesion indices. Those indices that demonstrated significant correlations were then included in a stepwise regression analysis to examine how well the indices predicted the variance in the human ratings. For the regression analysis, we used training and test sets to assess the generalizability of the regression model to an outside corpus. We divided the corpus of 171 essays into training and test sets following a 67/33 split (Witten et al., 2011). Those variables that demonstrated significant correlations with the human ratings were used as predictors in a regression analysis using the essays in the training set only. The model from this regression analysis was then applied to the held back essays in the test set to predict their ratings.

For each analysis, we also control for multicollinearity by examining correlations between indices and ensuring that the indices were not strongly related (i.e., $r < 0.70$). In addition, we control for overfitting in the models by ensuring a 15/1 item to predictor ratio. Controlling for statistical assumptions such as normality and multicollinearity allow us to use only variables that contribute uniquely to the models and to verify that the findings of the analysis are not the result of random noise in the data.

3. Results

3.1. Longitudinal growth human scores

A repeated measures ANOVA revealed a significant effect for time (initial, middle, and final essay) on total score, $F(2, 112) = 27.897$, $p < 0.001$, $\eta_p^2 = 0.333$. A significant linear trend, $F(1, 56) = 47.378$, $p < 0.001$, $\eta_p^2 = 0.458$, confirmed the expectation that time led to significant improvements in the total essay scores (see Table 4). Follow-up pairwise comparisons were conducted to evaluate differences among the means. The scores for the initial essays were significantly lower than those for the middle ($p < 0.001$) and final essays ($p < 0.001$); however, the difference between middle and final essays was not significant.

A repeated measures ANOVA revealed a significant effect for time (initial, middle, and final essay) on organization score, $F(2, 112) = 24.011$, $p < 0.001$, $\eta_p^2 = 0.300$. A significant linear trend, $F(1, 56) = 38.437$, $p < 0.001$, $\eta_p^2 = 0.407$, confirmed the expectation that time spent in a writing class led to significant improvements in the essay organization scores (see Table 4). Follow-up pairwise comparisons were conducted to evaluate differences among the means. The organization scores for the initial essays were significantly lower than for the middle essays ($p < 0.001$) and final essays ($p < 0.001$); however, the difference between middle and final essays was not significant.

Table 4
Mean total and organization scores for initial, middle, and final essays written across the semester.

Essay	Total score	Organization score
Initial	49.40 (10.29)	9.54 (2.07)
Middle	55.37 (7.99)	10.92 (1.75)
Final	57.11 (8.31)	11.41 (1.78)

Note: Standard deviation in parentheses.

Table 5
Mean cohesion index scores for initial, middle, and final essays written across the semester.

Index	Initial essay mean (SD)	Middle essay mean (SD)	Final essay mean (SD)	F	p	η^2
Adjacent overlap two paragraphs (nouns)	2.57 (1.54)	3.64 (2.20)	4.01 (2.03)	18.13	0.001	0.25
Content lemma TTR	0.59 (0.07)	0.56 (0.07)	0.55 (0.07)	14.97	0.001	0.21
LSA between all sentences and paragraphs	0.24 (0.11)	0.27 (0.09)	0.30 (0.07)	14.42	0.001	0.21
Incidence all positive connectives	0.16 (0.03)	0.18 (0.03)	0.18 (0.03)	14.17	0.001	0.20
Adjacent overlap two sentences (all words)	3.78 (1.51)	4.66 (1.68)	4.64 (1.75)	12.98	0.001	0.19
LSA initial to final paragraphs	0.42 (0.20)	0.48 (0.19)	0.53 (0.15)	11.54	0.001	0.17
Adjacent overlap two sentences (function words)	0.05 (0.02)	0.06 (0.02)	0.06 (0.02)	9.54	0.003	0.15
Bigram TTR	0.89 (0.04)	0.87 (0.05)	0.87 (0.04)	9.43	0.003	0.14
Adjacent overlap three sentences (function words)	0.10 (0.05)	0.12 (0.04)	0.12 (0.037)	9.19	0.004	0.14
Synonym overlap paragraphs (nouns)	2.95 (2.59)	3.95 (3.11)	4.76 (3.60)	9.67	0.003	0.15
Incidence all positive causal connectives	0.10 (0.03)	0.11 (0.03)	0.11 (0.03)	9.00	0.004	0.14
Syntactic similarity between sentences	0.14 (0.04)	0.13 (0.03)	0.13 (0.03)	6.89	0.011	0.11
Verb overlap	0.55 (0.14)	0.60 (0.11)	0.61 (0.12)	5.84	0.019	0.09

3.2. Longitudinal growth automated indices

Repeated measures ANOVAs were then conducted on the selected Coh-Metrix and TAACO indices to examine if significant differences in cohesive devices existed between initial, middle, and final essays. Of the 30 indices that were normally distributed, 13 demonstrated significant linear trends between the initial, middle, and final essays (see Table 5 for details). Of these variables, three were classified as global cohesion. These variables were related to lexical overlap (*adjacent overlap two paragraphs [nouns]*), semantic similarity *(LSA initial to final paragraphs)*, and synonym overlap *(synonym overlap paragraphs [nouns])*. Six variable were classified as local cohesion. These variables were related to connectives (*all positive connectives, incidence all positive causal connectives)* and lexical overlap *(adjacent overlap two sentences [all words], adjacent overlap two sentences [function words], adjacent overlap three sentences [function words], verb overlap sentences).* Four variables were classified as text cohesion. The variables were related to lexical diversity (*bigram TTR, content lemma TTR)*, semantic similarity *(LSA between all sentences and paragraphs)*, and syntactic cohesion *(syntactic similarity)*[3]. Pairwise comparisons between initial, middle, and final essays showed that, in all cases, final essays demonstrated significantly higher levels of cohesive devices than initial essays and in most cases middle essays demonstrated significantly higher levels of cohesive devices than initial essays (with the exceptions of *LSA initial to final paragraphs* and *noun synonym overlap between paragraphs*). However, in all cases, there were no significant differences between middle and final essays in terms of the production of cohesive devices.

3.2.1. Confirmatory support vector machine model

The 13 significant indices above were then used in a confirmatory SVM learning model to predict whether the essays were written at the beginning or end the semester. We did not focus on distinguishing middle semester essays because the pairwise comparisons reported no significant differences between middle and final essays. The SVM learning model using the 13 significant syntactic indices and 10-CV correctly allocated 81 of the 114 essays in the total set, χ^2 (df = 1, n = 114) = 20.217, $p < 0.001$, for an accuracy of 71.05% (the chance level for this analysis is 50%). The results from the SVM learning model are reported in the confusion matrix presented in Table 6. The measure of agreement between the actual category of the essay (i.e., beginning and end of semester) and that assigned by the model produced a Cohen's Kappa of 0.421, demonstrating moderate agreement.

3.3. Regression analyses: Organization

3.3.1. Correlations training set

Correlations were conducted between the cohesion indices and the human ratings of organization for the 119 (of the 171) essays selected for the training set by SPSS. Thirty Coh-Metrix and TAACO indices demonstrated significant correlations and reported at least a significant and small effect size ($r > 0.10$) with the human ratings while not demonstrating multicollinearity with one another (see Table 7).

3.3.2. Regression analysis training set

A stepwise regression analysis using these indices as the independent variables to predict the human ratings of organization yielded a significant model, F (4, 115) = 14.091, $p < 0.001$, $r = 0.574$, $r^2 = 0.329$. Three cohesion indices were included as significant predictors of the human ratings: *adjacent overlap two paragraphs (function words), function word TTR,*

[3] In these cases, the incidences of the cohesion features are not calculated at the sentence or paragraph level, but rather at the text level. Thus, depending on where the features occur, they could relate to either local and/or global cohesion.

Table 6
Confusion matrix for support vector machine algorithm.

		Initial essay	Final essay
LOOCV set	Initial essay	**41**	16
	Final essay	17	**40**

Table 7
Pearson correlations: Cohesion indices to organization scores.

Index	Cohesion type	*r*
Adjacent overlap two paragraphs (function words)	Global	0.54**
Adjacent overlap two paragraphs lemma (verbs)	Global	0.42**
Function word TTR	Text	−0.42**
Adjacent overlap two paragraphs (nouns)	Global	0.41**
Adjacent overlap two sentences (function words)	Local	0.37**
LSA initial to final paragraph	Global	0.29**
Adjacent overlap one paragraph (adverbs)	Global	0.29**
Causal links, verbs, and particles	Text	−0.26**
Adjacent overlap two sentences (all words binary)	Local	0.28**
Synonym overlap paragraphs (verbs)	Global	0.27**
Verb overlap	Text	0.25**
Incidence all positive connectives	Local	0.24**
Incidence all positive causal connectives	Local	0.21**
Syntactic similarity between sentences	Local	−0.20**
Adjacent overlap two sentences (verbs)	Local	0.20**
Synonym overlap paragraphs (nouns)	Global	0.19*
Adjacent overlap one paragraph (adjectives)	Global	0.19*
Adjacent overlap one sentence (pronouns)	Local	−0.18*
Incidence of conjunctions	Local	0.18*
Incidence of coordinating conjuncts	Local	−0.16*
Incidence of all causal connectives	Text	−0.16*
LSA initial to middle paragraph	Global	−0.16*
Lexical diversity: D	Text	0.16*
Incidence of conjuncts	Local	0.16*
Synonym overlap sentence (nouns)	Local	0.14*
Synonym overlap sentence (verbs)	Local	0.14*
Incidence of 'and'	Local	0.13*
Incidence of positive intentional connectives	Local	−0.12*
Bigram TTR	Text	−0.11*

* $p < 0.050$.
** $p < 0.010$.

incidence of coordinating conjuncts, and *adjacent overlap one sentence (pronouns)*. The first (*adjacent overlap two paragraphs (function words)* is a lexical overlap index related to global cohesion while the second is a lexical diversity index related to text cohesion *(function word TTR*) and the other two indices are connective and lexical overlap indices related to local cohesion. The model demonstrated that the four indices explained 33% of the variance in the human ratings of organization for the essays in the training set (see Table 8 for additional information).

3.3.3. Regression analysis test set

We used the model reported for the training set to predict the human ratings for organization in the test set. To determine the predictive power of the three variables retained in the regression model, we computed an estimated rating for each essay in the test set using the B weights and the constant from the reported training set regression model. A Pearson's correlation was then conducted between the estimated rating and the actual rating of each of the essays in the test set. This correlation together with its r^2 was then calculated to determine the predictive accuracy of the training set regression model on the independent data set.

The regression model, when applied to the test set, reported $r = 0.603$, $r^2 = 0.364$. The results from the test set model demonstrated that the combination of the three cohesion indices accounted for 36% of the variance in the organization ratings of the essays in the test set.

3.4. Regression analyses: Combined ratings

3.4.1. Correlations training set

Correlations were conducted between the cohesion indices and the combined human ratings of the 118 (of the 171) essays selected for the training set by SPSS. Thirty-two Coh-Metrix and TAACO indices of cohesion demonstrated significant

Table 8
Linear regression analysis to predict organization scores: training set.

Entry	Variable Added	*r*	r^2	*B*	SE	*B*
Entry 1	Adjacent overlap two paragraphs (function words)	0.49	0.21	0.14	0.04	0.32
Entry 2	Function word TTR	0.51	0.27	−11.23	3.69	−0.27
Entry 3	Incidence of coordinating conjuncts	0.55	0.30	−49.75	23.59	−0.17
Entry 4	Adjacent overlap one sentence (pronouns)	0.57	0.33	−14.44	7.05	−0.16

Notes: Estimated Constant Term is 13.676; *B* is unstandardized Beta; SE is standard error; B is standardized Beta.

correlations and reported a significant and at least a small effect size ($r > 0.010$) with the human ratings while not demonstrating multicollinearity with one another (see Table 9).

3.4.2. Regression analysis training set

A stepwise regression analysis using these indices as the independent variables to predict the combined human ratings yielded a significant model, $F\ (4,\ 114) = 53.781$, $p < 0.001$, $r = 0.666$, $r^2 = 0.443$. Four cohesion indices were included as significant predictors of the essay ratings: *adjacent overlap two paragraphs* (*function words*), *adjacent overlap two sentences* (*function words*), *adjacent overlap two paragraphs* (*pronouns binary*) and *pronoun to noun ratio*. Of these, two were lexical overlap indices related to global cohesion (*adjacent overlap two paragraph* [*function words*], *adjacent overlap two paragraphs*[*pronouns binary*]), one was a lexical overlap index related to local cohesion (*adjacent overlap two sentences* [*function words*]), and one was a givenness index related to text cohesion (*pronoun to noun ratio*). The model demonstrated that the four indices explained 44% of the variance in the combined human ratings of the essays in the training set (see Table 10 for additional information)

3.4.3. Regression analysis test set

The regression model, when applied to the test set, reported $r = 0.648$, $r^2 = 0.420$. The results from the test set model demonstrated that the combination of the four cohesion indices accounted for 42% of the variance in the combined ratings of the essays in the test set.

Table 9
Pearson correlations: Cohesion indices to combined scores.

Index	Cohesion type	*r*
Adjacent overlap two paragraphs (function words)	Global	0.59**
Lemma function word TTR	Text	−0.45**
Adjacent overlap two paragraphs (nouns)	Global	0.44**
Adjacent overlap two paragraphs (verbs)	Global	0.44**
Adjacent overlap two sentences (function words)	Local	0.42**
Incidence positive connectives	Local	0.33**
Synonym overlap paragraphs (verbs)	Global	0.33**
Adjacent overlap two sentences (all words)	Text	0.32**
Incidence positive causal connectives	Local	0.30**
Adjacent overlap two paragraphs (pronouns binary)	Global	0.28**
Causal links, verbs, and particles	Text	−0.26**
Syntactic similarity between sentences	Local	−0.25**
LSA initial to final paragraph	Global	−0.25**
Lexical diversity: D	Text	0.25**
Verb overlap	Text	0.24**
Incidence of conjunctions	Local	0.24**
LSA initial to middle paragraph	Global	0.23**
Synonym overlap paragraphs (nouns)	Global	0.22**
Number of motional prepositions	Text	0.21**
Synonym overlap sentences (verbs)	Local	0.19**
Incidence of 'and'	Local	0.17*
Incidence casual connectives	Text	−0.16*
Aspect repetition	Text	−0.15*
Incidence intentional connectives	Local	−0.15*
Pronoun to noun ratio	Text	−0.15*
Synonym overlap sentences (nouns)	Local	0.14*
Incidence of coordinating conjuncts	Local	−0.14*
Incidence of demonstratives	Text	0.13*
Incidence of conjuncts	Local	0.11*
Incidence of simple subordinators	Local	0.11*
Incidence negative connectives	Local	−0.11*

* $p < 0.050$.
** $p < 0.010$.

Table 10
Linear regression analysis to predict combined scores: training set.

Entry	Variable added	*r*	r^2	*B*	SE	*B*
Entry 1	Adjacent overlap two paragraphs (function words)	0.56	0.32	0.85	0.20	0.38
Entry 2	Adjacent overlap two sentences (function words)	0.62	0.39	67.51	19.79	0.28
Entry 3	Adjacent overlap two paragraphs (pronouns binary)	0.64	0.41	9.37	3.52	0.23
Entry 4	Pronoun to noun ratio	0.67	0.44	−6.61	2.67	−0.19

Notes: Estimated Constant Term is 37.343; *B* is unstandardized Beta; SE is standard error; B is standardized Beta.

4. Discussion

This analysis has demonstrated growth in the use of cohesive devices in L2 writers over the course of a semester. From the beginning of the semester until the end of the semester, L2 writers in this study generally wrote essays that demonstrated greater local, global, and text cohesion. The results of the study also indicate that indices of cohesion are predictors of human judgments of text organization and overall essay quality for L2 writing. However, cohesion patterns between the longitudinal analysis and the human judgments of quality show few similarities indicating a potential mismatch between cohesion growth and assessments of proficiency.

In terms of predicting human judgments of L2 writing organization, two indices that measure function word cohesion were predictive. The first index was a positive global predictor (adjacent overlap between paragraphs: function words) and the second measured text cohesion (function word TTR). The repetition of function words (i.e., the TTR index) was a negative predictor of essay quality. In addition to these two indices, two indices of local cohesion were also negative predictors of organization scores (incidence of coordinating conjunctions and sentence overlap of pronouns). For overall writing quality, four cohesive devices were predictive of human judgments. Two were global in nature (adjacent overlap between paragraphs for both function words and nouns), one was local in nature (adjacent overlap between sentences for function words), and the last measured text cohesion (pronoun to noun ratio). Only the last index was a negative predictor of human judgments of essay quality.

In reference to L2 cohesion development, this study demonstrated that a number of cohesive devices showed growth in predicted directions from the first to the final essay written within a semester. The strongest growth, as indicated by the effect size, was an increase in noun overlap between paragraphs, indicating an increase in global cohesion across essays. The next strongest variable, content lemma TTR demonstrated that writers increasingly repeated content words across a text (i.e., text cohesion). Two other indices supported the notion that writers increased cohesion at the text level by increasing the semantic similarity between all sentences and paragraphs (i.e., LSA between all sentences and paragraphs) and increasing the repetitions of bigrams across a text (i.e., a lower bigram TTR score). The remaining variables that showed gains were generally evenly split between local and global cohesion indicating that L2 writers increasingly used terms that were more strongly related across paragraphs (LSA initial to final paragraphs, noun synonyms between paragraphs) and across sentences by producing essays with more connectives, less lexical diversity, and greater lexical overlap between sentences (for all words and for verbs). In contrast to these general findings, one cohesion index demonstrated decreasing use over time: syntactic cohesion. However, this index is related to syntactic complexity as well as cohesion in the sense that lower syntactic similarity between sentences can indicate a greater variety of syntactic structures (cf. Crossley & McNamara, 2014). Thus, within a semester, we see growth in a variety of cohesion features. Specifically, we see growth in word overlap in terms of the overlap of ideas and function words at the sentence, paragraph, and text level. We also see growth in the use of connectives with the L2 writers in this sample producing a greater number of connectives over time. Thus, like previous studies, these findings support the notion that L2 learners exhibit growth at the local, global, and text level (Crossley et al., 2010a, 2010b; Yang & Sun, 2012)

Interestingly, of the 30 indices queried, over half showed no significant differences among the initial, middle, and final essays. These included most indices related to situation models reported by Coh-Metrix. Situation models apply to deeper cohesion features related to a text's temporal, causal, and spatial features (along with the intentionality of a texts) related to a reader's understanding of the text. Discontinuities in any of these dimensions within the text can cause a break in textual cohesion and make a text more difficult to process (Zwaan et al., 1995; Zwaan & Radvansky, 1998). However, because the texts sampled in this study are descriptive in nature and written by L2 writers, it may be the case that such links are unnecessary given the genre and because the writers may not have yet developed the skills to create such links in their writing. In addition to situation model indices, most connective indices did not show significant growth patterns indicating that the L2 writers in this study did not show development in terms of explicit cohesion links. However, they did demonstrate development in their use of more implicit cohesive devices such as lexical and semantic overlap between sentences and paragraphs.

We did not find similar patterns for cohesive devices when analyzing the indices that were most predictive of human judgments of organization and combined writing scores. The strongest predictors of judgments of text organization were calculated using function words and connectives (i.e., overlap of function words between paragraphs, function word TTR, incidence of coordinating conjuncts, and adjacent overlap of pronouns), explaining 36% of the variance of the human scores.

None of these indices demonstrated growth patterns in the longitudinal analysis (and the longitudinal indices that showed growth were mainly calculated using content words: nouns and verbs). Of the 13 cohesion variables in the longitudinal analysis, eight variables did show significant correlations with human judgments of text organization but none of these indices were included in the regression model. The indices that loaded in the combined writing score regression model were also all related to function words. Unlike the organization score regression, one local cohesion index that demonstrated growth in the longitudinal analysis was included in the combined writing score regression analysis (adjacent overlap two sentences[function words]). This index explained 7% of the variance for the human judgments in the training set. The other three variables, which did not show growth patterns in the longitudinal study, related to global cohesion and text cohesion (i.e., adjacent overlap two paragraphs [function words], adjacent overlap two paragraphs [pronouns binary], and pronoun to noun ratio) explained about 37% of the variance in the training set. Combined, the variables in the regression model explained 42% of the variance for the human scores in the test set. Of the 13 cohesion variables that showed significant growth in the longitudinal analysis, eight additional variables showed significant correlations with human judgments of combined scores, but seven of these variables were not used as predictors in the regression analysis.

The differences between the longitudinal study and the human judgments bear some discussion. The longitudinal analysis generally indicates that the use of cohesion features related to content words develops in learners more so than cohesive features related to function words. For instance, the two variables that reported the strongest effect sizes over time were based on content word overlap between paragraphs and across the text. A number of other variables that were predictive of longitudinal growth were also based on content words or a combination of content and function words. Only two variables were based on connectives and only two were uniquely based on function words. This differs from predictors of human judgments, which were mostly based solely on function words or connectives (in one case). It is possible, then, that human raters attend to more functional elements of overlap in text segments (e.g., determiners, conjunctions, prepositions, and pronouns) and that a writers' proficiency is associated with their ability to create these links in texts. This finding may be due to different expectations between L1 and L2 writing. In contrast to L2 writing, research into human ratings of L1 writing show that global cohesion markers focused on content word overlap are generally the most predictive features of rater quality. Knowing that the purpose of L2 writing is often not for content (as it is with L1 writers), but rather linguistic sophistication (Weigle, 2002), it make sense that L2 ratings are more strongly predicted by global overlap of function words related to text organization and syntax.

An additional finding from this study that bears examination is the number of positive correlations that were reported between local cohesive devices and text organization and combined essay scores. Such a finding counters many recent investigations into L2 writing (Crossley, Kyle, Allen, Guo, & McNamara, 2014a; Crossley & McNamara, 2012; Guo et al., 2013). One possible reason for these differences could be genre expectations. The majority of these previous studies have examined L2 writing proficiency using corpora of argumentative essays. These essays are based on independent tasks that are persuasive in nature and do not require domain or topic-specific knowledge. These tasks differ from the descriptive writing prompts found in this corpus. Thus, it is possible that descriptive writing prompts require more local and global cohesion as compared to independent prompts. Such differences have been reported between independent and integrated essays (i.e., essays that required the integration of source materials) with integrated essays demonstrating positive correlations with local cohesive devices (Guo et al., 2013). Second, many of the independent writing samples that have been analyzed in the past have been taken from standardized tests such as the Test of English as a Foreign Language (TOEFL). It may be the case that raters of standardized tests, which require more sophisticated writing approaches, may favor complexity over cohesion (Guo et al., 2013).

Differences also may exist between raters who are judging L1 writing and raters who are judging L2 writing. For instance, the majority of investigations into L1 writing show that human scores for text cohesion (Crossley & McNamara, 2010, 2011) and overall essay quality (Crossley & McNamara, 2010, 2011; Crossley et al., 2011; McNamara et al., 2013; McNamara, Crossley, Roscoe, Allen, & Dai, 2015) negatively correlate with indices of local cohesion, but positively correlate with indices of global cohesion in a manner dissimilar to that reported for L2 writers. Thus, it is possible that rater bias in terms of the first language of the writer plays a role in the assigning of essay quality scores. Support for such an argument may lie in L1 writing studies, which show that, at an early stage, L1 writers' use of both local and global cohesion is associated with higher essay quality. However, once young L1 writers reach a level at which they can coordinate sentences, they then begin to work toward composing at the global level by developing coherence between paragraphs (Bereiter & Scardamalia, 1987; Hayes & Flower, 1980). These global cohesive devices become a marker of quality for L1 writers, while local cohesive devices begin to be associated with poor writing quality (Berninger et al., 1996). Previous studies have also shown that L1 ratings of coherence are best explained by global and not local cohesion (Crossley & McNamara, 2011).

5. Conclusion

This study has provided evidence for using computational tools to examine growth patterns in the use of cohesive devices by L2 writers. An important element of this study was linking such growth to human ratings of text organization and combined writing quality. We found that, in most cases, cohesion cues that demonstrated growth patterns in L2 writers were not the same features that predicted judgments of proficiency in regression models. Such findings have important implications for both language testing and teaching. From a testing perspective, the results provide evidence for the types of cohesion devices that are predictive of essay quality providing some indication of what elements inform assessors' decisions.

Since these features are already automated, they could also be used to develop automatic essay scoring systems. From a teaching perspective, the results provide some clues to which cohesion elements show growth over a semester-long period. Knowing which cohesion elements demonstrate growth could help better inform teachers about the possible trajectories of their students and potentially allow them to better pinpoint instruction and intervention to target specific areas of cohesion development.

As always, some caution is warranted in interpreting the results. In this study, we have focused on a rather narrow element of the linguistic profile of writing: cohesion. While cohesion is an important element of writing (as seen in the reported effect sizes), it does not operate in isolation. Future studies may consider how cohesion interacts with other linguistic elements such as the lexicon and syntax in explaining growth and predicting writing quality. In addition, while we have measured a number of cohesive devices, there may be cohesive devices we did not calculate and some of our calculations may only tap into cohesion, but not fully measure it. Regarding the sample population, we only examined a small sample of ESL writers over a short span of time. Additionally, this population was of a limited range of proficiency levels. Future studies would benefit from a larger sample size of a greater range of proficiency levels observed over a longer period of time (e.g., a year-long program). Follow-up studies may also benefit from comparing instructed versus uninstructed learners and ESL to EFL learners. Lastly, this study focused on timed, descriptive writing. Future studies should consider assessing proficiency using a variety of different speaking and writing tasks to test the effects of genre and task. Such methodological changes would allow for falsification studies that could provide additional evidence in relation to the use of cohesive devices in L2 writing and their effects on human judgments of proficiency.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We also thank Charlene Polio and Jeff Connor-Linton for organizing the 2012 Georgetown University Round Table focusing on the data used in this study.

Appendix A

A.1. Set A

Describe your current home.
Describe a place you have visited recently.

A.2. Set B

Describe the campus of MSU.
Describe a good or bad teacher that you have had.

A.3. Set C

Describe your family.
Describe a good friend of yours.

A.4. Set D

Describe a school that you have attended.
Describe a problem that the United States or some other country is facing.

References

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written communication*. Hillsdale, NJ: Lawrence Erlbaum.
Berninger, V. W., Fuller, F., & Whitaker, D. (1996). A process model of writing development across the life span. *Educational Psychology Review, 8*, 193–218.
Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26*, 28–41.
Bestgen, Y., Lories, G., & Thewissen, J. (2010). Using latent semantic analysis to measure coherence in essays by foreign language learners? In S. Bolasco, I. Chiari, & L. Giuliano (Eds.), *Proceedings of the 10th international conference on statistical analysis of textual data (JADT)* (pp. 385–395). Rome: LED.
Bulte, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42–65.
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*, 257–269.
Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System, 31*, 471–484.
Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing, 26*, 1–9.
Cox, B. E., Shanahan, T., & Sulzby, E. (1990). Good and poor elementary readers' use of cohesion in writing. *Reading Research Quarterly, 25*, 47–65.

Crossley, S. A., Kyle, K., Allen, L. K., Guo, 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).. Retrieved from ⟨http://www.journalofwritingassessment.org/archives.php?issue=17⟩.
Crossley, S. A., Kyle, K., & McNamara, D. S. (in press). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*. doi: 10.3758/s13428-015-0651-7.
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*, 115–135.
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*, 66–79.
Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th international conference on artificial intelligence in education* (pp. 438–440). New York, NY: Springer.
Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*, 573–605.
Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010b). The development of semantic relations in second language speakers: A case for Latent Semantic Analysis. *Vigo International Journal of Applied Linguistics, 7*, 55–74.
Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010c). The role of lexical cohesive devices in triggering negotiations for meaning. *Issues in Applied Linguistics, 18*(1), 55–80.
Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*, 561–580.
Crossley, S. A., Yang, H. S., & McNamara, D. S. (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language, 26*(1), 92–113.
Crowhurst, M. (1987). Cohesion in argument and narration at three grade levels. *Research in the Teaching of English, 21*, 185–201.
Dufty, D. F., Graesser, A. C., Lightman, E., Crossley, S. A., & McNamara, D. S. (2006). An algorithm for detecting spatial cohesion in text. *Paper presented at the 16th annual meeting of the society for text and discourse*.
Englert, C. S., & Hiebert, E. H. (1984). Children's developing awareness of text structures in expository materials. *Journal of Educational Psychology, 76*, 65–74.
Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring for cohesive devices. In J., Oller, Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 177–181). Rowley, MA: Newbury House.
Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
Fitzgerald, J., & Spiegel, D. L. (1986). Textual cohesion and coherence in children's writing. *Research in the Teaching of English, 20*, 263–280.
Foltz, P. W. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 167–184). Mahwah, NJ: Lawrence Erlbaum.
Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing, 26*, 80–95.
Gernsbacher, M. A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum.
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, & Computers, 36*, 193–202.
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.
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
Haswell, R. H. (1986). *Change in undergraduate and post-graduate writing performance: Quantified findings (ERIC No. ED269780)*. ERIC Institute of Education Sciences Retrieved from ⟨http://eric.ed.gov/?id=ED269780⟩.
Haswell, R. H. (1990). *Change in undergraduate and post-graduate writing performance (part 2): Problems in interpretation (ERIC No. ED323537)*. ERIC Institute of Education Sciences Retrieved from ⟨http://eric.ed.gov/?id=ED323537⟩.
Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17*, 307–352.
Hayes, J., & Flower, L. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum.
Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language, 25*, 282–306.
Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System, 19*, 459–465.
King, M. L., & Rentel, V. (1979). Toward a theory of early writing development. *Research in the Teaching of English, 13*, 243–253.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review, 104*, 211–240.
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284.
Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System, 33*, 623–636.
Louwerse, M. (2001). An analytic and cognitive parametrization of coherence relations. *Cognitive Linguistics, 12*, 291–316.
Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*.. Basingstoke, UK: Palgrave Macmillan.
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)*. Retrieved from ProQuest Dissertations & Theses Global. (Order No. 3199485) (Doctoral dissertation).
McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*, 381–392.
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.
McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *Text-Interdisciplinary Journal for the Study of Discourse, 2*(1–3), 113–140.
McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*, 57–86.
McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45*, 499–515.
McNamara, D. S., Crossley, S. A., Roscoe, R., Allen, L., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35–59.
McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, MA: Cambridge University Press.
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.
McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*, 292–330.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*, 235–244.
Myhill, D. A. (2008). Towards a linguistic model of sentence development in writing. *Language and Education, 22*, 271–288.
Neuner, J. L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English, 21*, 92–105.
O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43*, 121–152.
Rentel, V. M., King, M. L., Pettegrew, B., & Pappas, C. (1983). *A longitudinal study of coherence in children's written narratives*. Columbus, OH: Ohio State University, Research Foundation.
Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing, 18*, 187–201.
Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication, 32*, 189–204.
Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann.
Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education, 23*, 31–48.
Yde, P., & Spoelders, M. (1985). Text cohesion: An exploratory study with beginning writers. *Applied Psycholinguistics, 6*, 407–415.
Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation-model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 386–397.
Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.

Scott Crossley is an Associate Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, and second language acquisition. He has published articles in second lexical acquisition, second language writing, second language reading, discourse processing, language assessment, intelligent tutoring systems, and text linguistics.