Journal of Research in Reading
UKLA
The United Kingdom Literacy Association

Journal of Research in Reading, ISSN 0141-0423 *DOI: 10.1111/j.1467-9817.2010.01449.x*
Volume 35, Issue 2, 2012, pp 115–135

Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication

Scott A. Crossley

Department of Applied Linguistics & ESL, Georgia State University, USA

Danielle S. McNamara

Institute for Intelligent Systems, University of Memphis, USA

This study addresses research gaps in predicting second language (L2) writing proficiency using linguistic features. Key to this analysis is the inclusion of linguistic measures at the surface, textbase and situation model level that assess text cohesion and linguistic sophistication. The results of this study demonstrate that five variables (lexical diversity, word frequency, word meaningfulness, aspect repetition and word familiarity) can be used to significantly predict L2 writing proficiency. The results demonstrate that L2 writers categorised as highly proficient do not produce essays that are more cohesive, but instead produce texts that are more linguistically sophisticated. These findings have important implications for L2 writing development and L2 writing pedagogy.

Over the last 30 years, a considerable amount of research has been conducted on the role of linguistic features in second language (L2) writing proficiency. This research has traditionally relied on surface and textbase measures such as text length, lexical diversity, word repetition and word frequency to distinguish differences between writing proficiency levels (e.g. Connor, 1990; Engber, 1995; Ferris, 1994; Frase, Faletti, Ginther & Grant, 1997; Jarvis, 2002; Jarvis, Grant, Bikowski & Ferris, 2003; Reid, 1986, 1990; Reppen, 1994). In contrast to these traditional measures of text, relatively little L2 writing research has been conducted using deeper-level linguistic measures (Engber, 1995) that tap into the underlying meaning and intentions of the discourse. Such measures assess linguistic features such as conceptual knowledge, causality, temporality and given/new information. This research gap has largely remained because of a paucity of accurate tools capable of adequately representing meaning and intention.

We address this research gap by using the computational tool Coh-Metrix (Graesser, McNamara, Louwerse & Cai, 2004) to examine the degree to which textual features can explain how the linguistic choices made by L2 writers relate to human judgements of writing proficiency. Unlike other computational tools, Coh-Metrix reports on a range of linguistic features at various levels of language, discourse, meaning and conceptual analysis. The indices reported by Coh-Metrix include surface- and textbase-level measures, as well as

deeper-level measures related to the situation model of the text (i.e. causality, temporality, inferencing and given/new information). In this study, we sample a range of theoretically motivated indices related to cohesion and linguistic sophistication, while controlling for an important non-linguistic variable: text length. Our goal is to examine whether L2 writers categorised at different proficiency levels produce writing samples that vary in terms of these linguistic features when text length is held constant.

Analysis of L2 writing

Of the four macro-skills related to communication, it has been argued that learning to write in a second language is far more challenging than learning to listen to, speak or read a foreign language (Bell & Burnaby, 1984; Bialystok, 1978; Brown & Yule, 1983; Nunan, 1989; White, 1981). Over the past 30 years, there have been a variety of methods and tools used to describe, distinguish and explain the writing processes unique to L2 learners. Many studies have examined effects of variables such as language background (Connor, 1996), the purpose of the writing, the writing medium (Biesenbach-Lucas, Meloni & Weasenforth, 2000), cultural expectations (Matsuda, 1997), topic and audience (Jarvis et al., 2003). In contrast, this study focuses on how differences in perceived writing proficiency are related to linguistic features present in the writers' texts. Our premise is that these features are tentacles to the writers' language abilities, which likely result from their exposure to the language and the amount of experience and practice they have in understanding and communicating in the second language (Dunkelblau, 1990; Kamel, 1989; Kubota, 1998). In this study, we focus specifically on language features related to cohesion (i.e. the use of connectives and word overlap) and linguistic sophistication (i.e. lexical difficulty, syntactic complexity). We selected language features related to cohesion and linguistic sophistication not only because they have been productive predictors of L2 writing proficiency in the past (e.g. Connor, 1990; Ferris, 1994; Frase et al., 1997; Grant & Ginther, 2000; Reid, 1986, 1990; Reppen, 1994; Silva, 1993), but also because recent developments in natural language processing allow us to consider deeper-level linguistic features related to cohesion and linguistic sophistication that were not available in previous studies.

Although research on the linguistic features of L2 writing has advanced recently, we still lack a coherent understanding of the linguistic features that characterise L2 writing (Jarvis et al., 2003). One reason that research in this area has lagged behind is related to the types of indices that have typically been available for exploration. Studies that have examined correlations between L2 essay scores and linguistic features have traditionally used surface code measures (Graesser, Millis & Zwaan, 1997). Surface code measures are those measures that assess word composition, lexical items, part of speech categories and syntactic composition at the surface level. In general, the studies that have used surface code measures have demonstrated that higher-rated essays contain more words (Carlson, Bridgeman, Camp & Waanders, 1985; Ferris, 1994; Frase et al., 1997; Reid, 1986, 1990), and use words with more letters or syllables (Frase et al., 1997; Grant & Ginther, 2000; Reid, 1986, 1990; Reppen, 1994). Syntactically, L2 essays that are rated as higher quality include more surface code measures such as subordination (Grant & Ginther, 2000) and instances of passive voice (Connor, 1990; Ferris, 1994; Grant & Ginther, 2000). Additionally, they contain more instances of nominalisations, prepositions (Connor, 1990), pronouns (Reid, 1992) and fewer present tense forms (Reppen, 1994).

Other studies have used measures that tap into the semantic textbase to evaluate the cohesive properties of L2 essays (Ferris, 1994; Silva, 1993). Unlike surface code measures, textbase indices identify explicit connections and referential links within text (Graesser et al., 1997). Such measures include indices of lexical diversity, word overlap and connectives. The findings from these studies have been contradictory. For instance, a number of studies have found that more proficient L2 writers use a more diverse range of words, and thus show greater lexical diversity (Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994). Greater lexical diversity signifies less word overlap and thus fewer referential links (McCarthy, 2005). Other studies have examined the use of more explicit cohesive devices such as connectives. Jin (2001), for example, examined the use of connectives in Chinese graduate students' writings. He found that all students, regardless of proficiency, use cohesive devices but advanced writers use these devices more often than do intermediate writers. Similarly, Connor (1990) found that higher-proficiency L2 writers use more connectives. Past research, then, demonstrates that L2 writers judged to be advanced sometimes produce text which is less cohesive when measured by word overlap, but at other times their writing is more cohesive as measured by their use of connectives.

Overall, studies using surface code and textbase measures to compare incidences of linguistics features in L2 essays to their respective essay scores demonstrate that linguistic variables related to cohesion and linguistic sophistication (i.e. lexical repetition, connectives, parts of speech, word length, lexical diversity and the use of passive voice) along with text length can be used to distinguish high-proficiency essays from low-proficiency essays. However, while these studies have made important contributions to our understanding of L2 writing proficiency, many of the studies have suffered from design weaknesses. Additionally, the reported findings, while statistically significant, have generally demonstrated low-to-moderate effect sizes (defined as Pearson's correlations $<.50$, Cohen, 1988) such as $r<.40$ in Engber (1995) and Jarvis et al. (2003) or, like Grant and Ginther (2000), have not reported effect sizes at all. In reference to design weaknesses, some studies have compromised statistical validity by potentially over-fitting the data (e.g. Jarvis et al., 2003). Over-fitting data are problematic because if too many variables are included in a statistical analysis, the model fits not only the signal of the predictors but also the unwanted noise. When over-fitting occurs, the training model fits the data well, but when the model is applied to new data, the results are likely to be poor because noise varies across data sets. Thus, the noise fit to the model in the original data set will not remain the same in a new data set. Additionally, most, if not all past studies, have failed to use training sets and test sets (e.g. Connor, 1990; Engber, 1995; Ferris, 1994; Frase et al., 1997; Jarvis et al., 2003; Reid, 1986, 1990). In these cases, the reported performance may fit the analysed corpus, but there is no test provided to indicate whether the linguistic variables will provide good predictors of performance on other corpora (Whitten & Frank, 2005).

While these past studies have strongly contributed to our knowledge of L2 writing, the reported results are not always extendable outside of the data analysed. As such, more research is needed to validate the role of linguistic features in characterising L2 essay quality. Additionally, other linguistic features need to be considered to investigate potentially better and more reliable predictors of essay quality. The use of more advanced computational tools to identify these features along with larger corpora (both of which were unavailable to past researchers and account for the reported design weaknesses) should likely facilitate this effort.

Computational tools and text evaluation

While still far from the norm, the use of computational tools in the examination of L2 writing is steadily growing. Past studies using computational tools have included the use of STYLEFILES (Reid, 1992) and computerised tagging systems (Grant & Ginther, 2000; Jarvis et al., 2003). More recently, L2 writing researchers have begun to take advantage of computational tools that provide more sophisticated linguistic indices, such as Coh-Metrix (Crossley & McNamara, 2009; McCarthy et al., 2007).

Coh-Metrix is an advanced computational tool that measures cohesion and linguistic sophistication at various levels of language, discourse and conceptual analysis (McNamara, Crossley & McCarthy, 2010). The tool was constructed to investigate various measures of text and language comprehension that augment surface components of language by exploring deeper, more global attributes of language. The tool is informed by various disciplines such as discourse psychology, computational linguistics, corpus linguistics, information extraction and information retrieval. As such, Coh-Metrix integrates lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters and other components common in computational linguistics (Jurafsky & Martin, 2008). Coh-Metrix indices have been validated in numerous writing studies (see Crossley & McNamara, 2009, for an overview).

With these resources, Coh-Metrix can analyse text using measures of both the surface code and textbase. Notably, many of the surface code and textbase measures provided by Coh-Metrix reports have not been used to analyse L2 writing in prior research. These Coh-Metrix measures include depth of knowledge lexical indices, word overlap indices and syntactic complexity indices (Graesser et al., 2004). More importantly, Coh-Metrix provides a selection of indices related to textual features that are likely to influence a reader's deeper understanding of the text, called the situation model (Kintsch, 1998). Some researchers have proposed that the reader's situation model is influenced by various dimensions of the text related to textual coherence (i.e. causality, temporality, intentionality and protagonists; Zwaan, Magliano & Graesser, 1995; Zwaan & Radvansky, 1998). Discontinuities in any of these dimensions within the text can cause a break in textual cohesion, and thus affect the coherence of the reader's situation model.

Current study

Our goal is to take advantage of the more complete range of indices provided by Coh-Metrix to examine potentially better indicators of human judgements of L2 writing proficiency. We hypothesise that linguistic features related to cohesion and linguistic sophistication will provide strong predictors of human judgements of writing proficiency. Our approach using a broader range of indices that are linked to the surface code, textbase and situation model, contrasts with prior L2 writing studies that have examined a limited number of linguistic features assessed solely by surface and textbase measures. In addition, we examine collections of linguistic features that are related to specific cognitive correlates such as cohesion (e.g. logical operators, lexical overlap, temporal cohesion, semantic co-referentiality, causality, connectives, lexical diversity) and linguistic sophistication (e.g. psycholinguistic word ratings, hypernymy, word frequency, syntactic complexity). The categorisations we use are based on theoretical conventions and are discussed below. The categories are not unambiguous and a few of the categories exhibit theoretical convergence. For instance, lexical diversity has tentacles to both

cohesion and linguistic sophistication. Nonetheless, our groupings are helpful in understanding the intended constructs.

We use the term cohesion to refer to the textual indications that coherent texts are built upon (Louwerse, 2004). Cohesion is critical to the understanding of how language functions and is premised on the notion that the linking of ideas allows for the creation of coherent discourse (Halliday & Hasan, 1976). In this study, we predict that higher-rated essays will contain more cohesive devices than lower-rated essays. This prediction is based on past studies (Connor, 1990; Jin, 2001; Witte & Faigley, 1981) that found that more proficient writers produced more cohesive devices than less proficient writers.

We use the term linguistic sophistication to refer to the production of infrequent and more complex linguistic features. Linguistic sophistication is important because it relates to the depth of linguistic knowledge by language learners and strongly correlates with language proficiency and academic achievement (Daller, van Hout & Treffers-Daller, 2003). In this study, we predict that more proficient writers will demonstrate greater linguistic sophistication than lower-proficiency writers, especially in relation to lexical difficulty. This prediction is based on past studies (e.g. Grant & Ginther, 2000; Reppen, 1994) that have shown that advanced writers use more difficult and varied linguistic items (in both word and syntactic choices).

Method

To accomplish our goal of determining the combined effects of surface code, textbase and situational model measures on essay evaluation, we used Coh-Metrix to analyse a corpus of scored essays that were controlled for text length. Unlike past studies, we divided the texts into *training* and *test* sets (Whitten & Frank, 2005). Using the training set data, we conducted correlations and linear regressions comparing the human ratings and the Coh-Metrix variables. The results of this analysis were later extended to the held back, independent test set data and finally to the complete corpus.

Corpus collection

We used essays written by graduating Hong Kong high school students for the Hong Kong Advanced Level Examination (HKALE). Milton (2000) initially used the corpus to examine writing proficiency at various levels for lexical and grammatical acquisition. The complete corpus, which Milton referred to as the HK 'UE' Examination Scripts corpus, consists of 1,200 essays. The essays were administered to senior high school students and were designed to assess students' ability to understand and use English at the college level. The writing examination lasted for 1 hour and 15 minutes. Participants were expected to write a 500-word essay and were allowed to choose from one of four prompts. The essays in the corpus were written in response to the following prompts: *discuss the popularity of comic books*, *discuss the wearing of brand named fashions*, *respond to a letter of complaint* and *write a letter welcoming an exchange student*. The essays were graded by groups of trained raters from the Hong Kong Examinations and Assessment Authority. The corpus includes six of the seven represented grade levels assigned to HKALE (the grade of *unclassifiable* was left out). Thus, the grades ranged from A to F (and included the grade E). The corpus comprised 200 essays from each grade level, for a total of 1,200 essays.

The corpus used in this study is a subsection of the HK 'UE' Examination Scripts corpus. We selected only the essays that had text lengths between 485 and 555 words. These essays provided us with the greatest number of texts (514) that did not demonstrate significant correlations between text length and the grades assigned by the raters. We controlled for text length effects because text length has historically been a strong predictor of essay scoring with most studies reporting that text length explains about 30% of the variance in human scores (Ferris, 1994; Frase et al., 1997). Additionally, when text length is combined with other variables, it generally washes out their predictive strength. Given that our interest lies in linguistic variables related to cohesion and linguistic sophistication, we selected a subsection of texts that did not demonstrate significant correlations between text length and human scoring. This selection allowed us to examine which linguistic variables influence human scoring when text length is held constant.

Variable selection

Coh-Metrix reports over 600 indices of linguistic features of text. All indices reported by Coh-Metrix are normalised for text length. For this study, we divided these indices into 12 conceptually similar *banks* related to cohesion and linguistic sophistication. To select the variables from the Coh-Metrix banks of indices (e.g. word frequency bank, syntactic complexity bank, lexical diversity bank), we followed Whitten and Frank (2005) and divided the corpus into two sets: a training set ($n = 344$) and a testing set ($n = 170$) based on a 67/33 split. The training set was used to select the linguistic variables. The test set was used to calculate the amount of variance that the selected variables explained in an independent corpus (Whitten & Frank, 2005). Such a method allowed us to predict accurately the performance of our model on an independent corpus. Because the selected essays were controlled for text length, there was not an even division of essays based on grade categorisation. The fewest essays were contained in the 'A' categorisation (63 essays). The categorisation with the most essays was the 'C' categorisation (94 essays). A description of the training and test set is located in Table 1.

The purpose of the training set was to identify which of the Coh-Metrix variables most highly correlated with the essay grades. Those variables that most highly correlated with the essay grades were used in a regression analysis to examine if the Coh-Metrix variables were predictive of human essay ratings. To avoid issues of collinearity (strong correlations between two or more variables), we conducted Pearson's correlations between the variables to ensure that none of the indices correlated at $r = > .70$ (Brace Kemp & Snelgar, 2006; Tabachnick & Fidell, 2001). The selected banks are discussed below in reference to their importance in text cohesion and linguistic sophistication. The banks are also separated based on whether the indices they report are based on the surface code, the textbase or the situation model.

Table 1. Number of essays contained in each grade classification.

Grade	Total number of essays	Essays in training set	Essays in test set
A	63	42	21
B	88	59	29
C	94	63	31
D	87	58	29
E	89	60	29
F	93	62	31

Surface code measures

Syntactic complexity. Syntactic complexity is measured by Coh-Metrix in three major ways. First, there is an index that calculates the mean number of words before the main verb with the understanding that more words before the main verb leads to more complex syntactic structure. Second, there is an index that measures the mean number of high-level constituents (sentences and embedded sentence constituents) per word with the understanding that more higher-level constituents per word leads to a more complex syntactic structure. Lastly, there is an index that assesses syntactic similarity by measuring the uniformity and consistency of the syntactic constructions in the text. This index not only looks at syntactic similarity at the phrasal level, but also takes account of the parts of speech involved. More uniform syntactic constructions result in less complex syntax that is easier for the reader to process. Sentences with difficult syntactic constructions include the use of embedded constituents and are often structurally dense, syntactically ambiguous or ungrammatical (Graesser et al., 2004). As a consequence, they are more difficult to process and comprehend (Perfetti, Landi & Oakhill, 2005).

Word frequency. Word frequency refers to metrics of how often particular words occur in the English language. Unlike past word frequency measures that simply look at bands of frequent words and generally only the first 2,000 words (Nation & Heatley, 1996), Coh-Metrix reports on a more sophisticated word frequency measure that is based on large corpora. The primary frequency count in Coh-Metrix comes from CELEX (Baayen, Piepenbrock & van Rijn, 1993), the database from the Centre for Lexical Information, which consists of frequencies taken from the early 1991 version of the COBUILD corpus, a 17.9 million-word corpus. Thus, Coh-Metrix reports on the incidence of word frequency for a majority of the words in English, not just the most common. From a cognitive perspective, frequent words are more quickly accessed by the writer and more quickly decoded by the reader (Perfetti, 1985; Rayner & Pollatsek, 1994). More proficient L2 writers have also been shown to use less frequent words (Frase et al., 1997; Grant & Ginther, 2000; Reid, 1986, 1990; Reppen, 1994).

Hypernymy and polysemy indices (WordNet). Coh-Metrix measures the ambiguity of a text by calculating its polysemy value, which refers to the number of meanings or senses within a word. Coh-Metrix measures the abstractness of a text by calculating its hypernymy value, which refers to the number of levels a word has in a conceptual, taxonomic hierarchy (from concrete to abstract). The number of meanings and the number of levels attributed to a word are measured in Coh-Metrix using WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross & Miller, 1990). For instance, on a hypernymic scale, *vehicle* would have more levels and thus be more abstract than *car*. Using polysemy, *bank* (*go to the bank*, *break the bank*, *bank on the Yankees winning*) would be more ambiguous than *lentil*, which has only one sense. Hypernymy and polysemy values also relate to the development of L2 lexical networks because hypernymy values can measure the growth of lexical connections between hierarchically related items, while polysemy values can measure the development of sense relations (Crossley, Salsbury & McNamara, 2009; in press).

Word information (MRC psycholinguistic database). Coh-Metrix calculates information at the lexical level on five psycholinguistic matrices: familiarity, concreteness,

imagability, meaningfulness and age of acquisition. All of these measures come from the MRC psycholinguistic database (Coltheart, 1981) and are based on the works of Paivio (1965), Toglia and Battig (1978) and Gilhooly and Logie (1980), who used human subjects to rate large collections of words for said psychological properties. Because most MRC measures are based on psycholinguistic experiments, the coverage of words differs among the measures (e.g. the database contains 4,825 words with imagery ratings and 4,920 with familiarity ratings). Many of these indices are important for L2 lexical networks and lexical difficulty. In relation to lexical networks, the MRC word meaningfulness score relates to how strongly words associate with other words, and how likely words are to prime or activate other words. In relation to lexical difficulty, MRC word familiarity, concreteness, imagability and age of acquisition scores measure lexical constructs such as word exposure (familiarity), word abstractness (concreteness), the evocation of mental and sensory images (imagability) and intuited order of lexical acquisition (age of acquisition). For a full review of these indices as found in Coh-Metrix refer to Salsbury, Crossley and McNamara (in press).

Textbase measures

Lexical overlap. Coh-Metrix considers four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap and content word overlap. Noun overlap measures how often a common noun of the same form is shared between two sentences. Argument overlap measures how often two sentences share nouns with common stems (including pronouns), while stem overlap measures how often a noun in one sentence shares a common stem with other word types in another sentence (not including pronouns). Content word overlap refers to how often content words are shared between sentences at binary and proportional intervals (including pronouns). Lexical overlap has been shown to aid in text comprehension (Douglas, 1981; Kintsch & van Dijk, 1978; Rashotte & Torgesen, 1985). Lexical overlap indices are also of interest because advanced L2 writers have been found to use a greater variety of lexical and referential cohesion devices, while lower-level writers use more overlap (Ferris, 1994).

Connectives. In Coh-Metrix, the density of connectives is assessed using two dimensions. The first dimension contrasts positive versus negative connectives, whereas the second dimension is associated with particular classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001). These connectives are associated with positive additive (*also*, *moreover*), negative additive (*however*, *but*), positive temporal (*after*, *before*), negative temporal (*until*) and causal (*because*, *so*) measures. Connectives play an important role in the creation of cohesive links between ideas and clauses (Crismore, Markkanen & Steffensen, 1993; Longo, 1994) and provide clues about text organisation (Van de Kopple, 1985).

Logical operators. The logical operators measured in Coh-Metrix include variants of *or*, *and*, *not* and *if-then* combinations, all of which have been shown to relate directly to the density and abstractness of a text and correlate to higher demands on working memory (Costerman & Fayol, 1997).

Lexical diversity. The traditional method for measuring lexical diversity (LD) is type–token ratio (TTR, Templin, 1957). TTR is the division of types (i.e. unique words

occurring in the text) by tokens (i.e. all instances of words), forming an index that ranges from 0 to 1, whereby a higher number indicates greater diversity. However, a major problem with LD indices is that while the number of tokens increases uniformly, the relative number of types steadily decreases. That is, every new word is a new token; however, after a relatively short amount of text, tokens tend to be repeated such that the number of types asymptotes. As a result, TTR indices are generally highly correlated to text length and are not reliable across a corpus of texts where the token counts differ markedly (McCarthy & Jarvis, 2007). To correct for the problem of text length in LD indices, a wide range of more sophisticated approaches to measuring vocabulary range have been developed. Those reported by Coh-Metrix include MTLD (McCarthy, 2005; McCarthy & Jarvis, in press) and *D* (Malvern, Richards, Chipere & Duran, 2004; McCarthy & Jarvis, 2007) values. LD is indicative of the range of vocabulary deployed by a speaker or writer. Greater LD is widely held to be indicative of greater linguistic skills (Avent & Austermann, 2003; Grela, 2002) and past studies have demonstrated that more proficient L2 writers produce texts with greater lexical diversity (Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994).

Situation model measures

Latent semantic analysis (LSA). Coh-Metrix tracks semantic coreferentiality using LSA, a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts. Unlike indices of lexical overlap, LSA measures semantic similarity between words, sentences and paragraphs. LSA uses a general form of factor analysis to condense a large corpus of texts down to 300–500 dimensions. These dimensions represent how often a word occurs within a document (defined at the sentence level, the paragraph level or in larger sections of texts) and each word, sentence or text is represented by a weighted vector (Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998). The relationships between the vectors form the basis for representing semantic similarity between words. For instance, the word *mouse* has a higher LSA value when compared with *cat* than to either *dog* or *house*.

In addition, Coh-Metrix assesses given/new information through LSA by measuring the proportion of new information each sentence provides. The given information is thought to be recoverable from the preceding discourse (Halliday, 1967) and does not require a memory search (Chafe, 1975). Given information is thus less taxing on a person's cognitive load. To compute the LSA given/new index, each sentence in the input text is represented by an LSA vector. Then the amount of new information a sentence provides is computed from the component of the corresponding sentence vector that is perpendicular to the space spanned by the previous sentence vectors. Similarly, the amount of given information of a sentence is the parallel component of the sentence vector to the span of the previous sentence vectors (Hempelmann, Dufty, McCarthy, Graesser, Cai & McNamara, 2005). LSA indices are important measures of cohesion because they can track the amount of semantic coreferentiality in a text (Crossley, Louwerse, McCarthy & McNamara, 2007). LSA is also indicative of the development of lexical networks by L2 learners (Crossley, Salsbury, McCarthy & McNamara, 2008).

Spatiality. Coh-Metrix determines spatial cohesion based on the work of Herskovits (1998) who suggested that there are two kinds of spatial information: location information and motion information. This theory is extended by representing motion

spatiality through motion verbs such as *run*, *drive* and *move* and location spatiality through location nouns such as *Alabama*, *house* and *store* (Dufty, Graesser, Lightman, Crossley & McNamara, 2006). In Coh-Metrix, classifications for both motion verbs and location nouns are taken from WordNet (Fellbaum, 1998). Coh-Metrix uses this information to produce a variety of indices related to spatiality. Key among these are the ratio of location and motion words, the number of locational nouns, the number of locational prepositions and the number of motion verbs. Spatial cohesion helps to construct a text and ensures that the situational model of the text (Kintsch & van Dijk, 1978; Zwaan et al., 1995) is well structured and clearly conveys text meaning.

Causal cohesion. Cues that help infer the causal relations in the text (i.e. causal cohesion) enhance situation model-level understanding (Kintsch & van Dijk, 1978; Zwaan et al., 1995). Causal cohesion is measured in Coh-Metrix by calculating the ratio of causal verbs to causal particles (Graesser et al., 2004). The incidence of causal verbs and causal particles in a text relates to the conveyance of causal content and causal cohesion. The causal verb count is based on the number of main causal verbs identified through WordNet (Fellbaum, 1998; Miller et al., 1990). These include verbs such as *kill, throw* and *pour*. The causal particle count is based on a defined set of causal particles such as *because*, *as a consequence of* and *as a result*. Causal cohesion is relevant to texts that depend on causal relations between events and actions (i.e. stories with an action plot or science texts with causal mechanisms) and is also relevant at the sentential level when there are causal relationships between sentences or clauses (Pearson, 1974–1975).

Temporality. Temporal cues help construct a more coherent situation model of a text. There are three principal measures in Coh-Metrix related to temporality: aspect repetition, tense repetition and the combination of aspect and tense repetition. Linguistic features related to tense and aspect are foundational for interpreting temporal coherence in texts (Duran, McCarthy, Graesser & McNamara, 2007). Tense helps to organise events along timelines and can affect the activation of information in working memory. Tense also relates lexical events to a certain point in time. Aspect, on the other hand, conveys the dynamics of the point itself such as whether the point is ongoing or completed (Klein, 1994). Aspect also helps maintain information in working memory (Magliano & Schleich, 2000). It is argued that more experienced writers repeat tense and aspect as a means of creating greater cohesion in their writing (Duran et al., 2007).

Results

Pearson's correlations training set

We selected the variables from each bank of Coh-Metrix indices that demonstrated the highest, significant Pearson's correlation when compared with the human scores of the L2 writers' essay. We selected multiple indices from the MRC database because the indices did not measure the similar constructs. The 14 selected variables and their banks along with their r values and p values are presented in Table 2, sorted by the strength of the correlation. Only one bank, syntactic complexity, did not contain a variable that correlated significantly to the essay scores. To control for over-fitting, we followed a conservative approach that allowed for one predictor per 20 items. With 344 essays in the

Table 2. Selected Coh-Metrix variables based on Pearson's correlations with essay grade.

Variable	Discourse level	Cognitive correlate	Bank	*r* value	*p* value
***D* Lexical Diversity**	Textbase	Linguistic sophistication/ cohesion	Lexical diversity	0.426	$<.001$
Word familiarity	Surface code	Linguistic sophistication	MRC database	-0.400	$<.001$
CELEX content word frequency	Surface code	Linguistic sophistication	Word frequency	-0.336	$<.001$
Content word overlap	Textbase	Cohesion	Word overlap	-0.279	$<.001$
LSA given/new	Situation model	Cohesion	Given/new	-0.265	$<.001$
Incidence of positive logical connectives	Textbase	Cohesion	Connectives	-0.227	$<.001$
Word concreteness	Surface code	Linguistic sophistication	MRC database	-0.209	$<.001$
Word imagability	Surface code	Linguistic sophistication	MRC database	-0.180	$<.001$
Word meaningfulness	Surface code	Linguistic sophistication	MRC database	-0.176	$<.001$
Aspect repetition	Situation model	Cohesion	Temporal cohesion	-0.163	$<.050$
LSA sentence to sentence	Situation model	Cohesion	Semantic similarity	-0.150	$<.001$
Number of motion verbs	Situation model	Cohesion	Spatial cohesion	0.124	$<.050$
Logical operators	Textbase	Cohesion	Logical operators	0.122	$<.050$
Verb hypernymy	Surface code	Linguistic sophistication	WordNet	0.121	$<.050$
Number of words per essay	Surface code	None	Text length	0.095	$>.050$

Notes: LSA given/new and word imagability were not included in the analysis to reduce collinearity; indices in bold remained significant predictors in the linear regression analysis.

training set, we could safely include all 14 of the variables in the regression analysis (Brace et al., 2006; Field, 2005; Tabachnick & Fidell, 2001).

Collinearity

Pearson's correlations demonstrated that the content word overlap variable was highly correlated ($>.70$) with the LSA given/new measure ($N = 344$, $r = -.741$, $p < .001$). Because the content word overlap measure had the highest correlation with essay scores between the two variables, it was retained in the analysis and the LSA given/new measure was dropped. Pearson's correlations also demonstrated a high correlation between word imagability and word concreteness ($N = 344$, $r = -.928$, $p < .001$). Because the word concreteness index had the highest correlation with essay scores, it was retained and the word imagability index was dropped. Thus, 12 variables were included in the analysis.

Multiple regression training set

A linear regression analysis was calculated for the 12 selected variables. These 12 variables were regressed onto the raters' evaluations for the 344 essays in the corpus. The variables were also checked for outliers and multicollinearity. Coefficients were checked for both variance inflation factors (VIF) values and tolerance. All VIF values were at about 1 and all tolerance levels were beyond the .2 threshold, indicating that the model data did not suffer from multicollinearity (Field, 2005).

The linear regression using the 12 variables yielded a significant model, $F(5, 338) = 28.278$, $p < .001$, $r = .543$, $r^2 = .295$. Five variables were significant predictors in the regression: *D* (lexical diversity), word familiarity, CELEX content word frequency, word meaningfulness and aspect repetition. Descriptive statistics for these five variables are provided in Table 3. The regression analysis demonstrates that the combination of the five variables accounts for 30% of the variance in the evaluation of the 344 essays examined in the training set (see Table 4 for additional information). Seven variables were not significant predictors: logical operators, motion verbs, verb hypernymy, word concreteness, incidence of logical positive connectors, content word overlap and LSA sentence to sentence. The latter variables were left out of the regression model; *t*-test information on these variables and the variables from the regression model as well as the amount of variance explained is presented in Table 5.

Test set model

To further support the results from the multiple regression conducted on the training set, we used the B weights and the constant from the training set multiple regression analysis to estimate how the model would function on an independent data set (the 170 evaluated essays held back in the test set). The model produced an estimated value for each essay in the test set. We then conducted a Pearson's correlation between the estimated score and the actual score. We used this correlation along with its r^2 to demonstrate the strength of the model on the independent data set. Descriptive statistics for the selected variables and the essay evaluations from the test set are provided in Table 6. The model for the test set yielded $r = .454$, $r^2 = .206$. The results from the test set model demonstrate that the combination of the five variables accounted for 21% of the variance in the evaluation of the 170 essays examined in the test set.

Total set model

To examine how the model from the training set predicted the variance in L2 essays scores for the entire corpus, we used the B weights and the constant from the training set multiple regression analysis on the entire data set (the 514 evaluated essays that were contained in both the training and the test set). Such a methodology allows us to test and validate the model yielded in the training set and determine with confidence that the training model reported was not the result of over-fitting. If the total set model is similar to the training set model, we can say with a higher degree of confidence that it is a reliable model (Whitten & Frank, 2005). We followed the same methodology for the total set model as for the test set model. Descriptive statistics for the selected variables and the essay evaluations from the total set are provided in Table 7. The model for the entire data set yielded $r = .509$, $r^2 = .259$. The results from the entire data set model demonstrate that the combination of the five variables accounts for 26% of the variance in the evaluation of the 514 essays that comprise our L2 writing corpus.

Discussion

This study provides evidence that surface code, textbase and situation model measures related to cohesion and linguistic sophistication at least partially characterise human judgments of proficiency in L2 writing. Understanding the function of such features and

Table 3. Training set statistics (*M*, *SD*) for Coh-Metrix indices as a function of grade.

Coh-Metrix index	Assigned grade/rating					
	A	B	C	D	E	F
Lexical diversity *D*	109.667 (30.793)	82.881 (22.465)	83.048 (20.882)	77.172 (21.600)	71.50 (22.934)	60.307 (20.789)
CELEX content word frequency	1.116 (0.221)	1.18 (0.212)	1.237 (0.205)	1.277 (0.228)	1.329 (0.232)	1.356 (0.230)
Word meaningfulness	615.588 (44.612)	633.497 (41.203)	637.823 (44.998)	637.413 (40.788)	644.928 (42.990)	643.936 (43.242)
Average of word familiarity	592.453 (2.649)	594.727 (2.668)	594.493 (2.489)	594.670 (2.754)	596.201 (2.807)	596.796 (2.968)
Aspect repetition	0.921 (0.077)	0.923 (0.087)	0.945 (0.056)	0.948 (0.062)	0.950 (0.074)	0.954 (0.063)

Table 4. Linear regression analysis findings to predict essay ratings: training set.

Entry	Variable added	Correlation	R^2	*B*	B	*SE*
Entry 1	*D* lexical diversity	0.427	0.180	0.011	0.170	0.004
Entry 2	CELEX content word frequency	0.466	0.213	− 2.387	0.337	0.430
Entry 3	Average of word meaningfulness	0.52	0.264	− 0.008	− 0.198	0.002
Entry 4	Average of word familiarity	0.534	0.276	− 2.590	− 0.111	1.094
Entry 5	Aspect repetition	0.543	0.285	− 0.077	− 0.14	0.035

Notes: *B* = unstandardised β; B = standardised β; *SE* = standard error.
Estimated constant term is 58.486.

Table 5. *t*-values, *p*-values and variance explained for training set variables.

Variable	*t*-value	*p*-value	R^2
D Lexical diversity	2.789	< .010	0.18
CELEX content word frequency	− 5.552	< .001	0.04
Average of word meaningfulness every word	− 3.132	< .010	0.05
Aspect repetition score	− 2.369	< .050	0.02
Average of word familiarity every word	− 2.217	< .050	0.01
Logical operators	1.899	> .050	0
Number of motion verbs	− 1.041	> .050	0
Verb hypernymy	1.035	> .050	0
Average of word concreteness	0.992	> .050	0
Incidence of positive logical connectives	− 0.978	> .050	0
Content word overlap	− 0.571	> .050	0
LSA sentence to sentence	− 0.008	> .050	0

how they relate to writing proficiency significantly impacts our understanding of the importance of linguistic features in essays by explaining the role text variables play in predicting writing proficiency. These findings can be used to inform writing pedagogy and provide models for computer-assisted language learning.

Unlike many past L2 studies that have examined lexical, grammatical and discourse features (Engber, 1995; Grant & Ginther, 2000; Jarvis et al., 2003), this study demonstrates large effect sizes (defined as Pearson's correlations $\geqslant .50$, Cohen, 1988) between the selected lexical features and the essay scores. Additionally, the computational tools and the size of the corpus employed in this study allow us to use more rigorous statistical methodology and control for issues such as over-fitting. Thus, we have confidence that this analysis provides reliable evidence that computational measures of linguistic features can be used to predict L2 writing proficiency. We are also confident that the findings are generalisable, at least for the language population surveyed. Our confidence rests in the use of both training and test sets.

The findings from this study also demonstrate the potential for surface code, textbase and situation model measures related to cohesion and linguistic sophistication to predict essay scores. The surface code and textbase variables examined in this study (*lexical diversity*, *word familiarity*, *word frequency*, *word meaningfulness*) accounted for the vast majority of the variance in the multiple regression model with the variable *D* (*lexical diversity*) accounting for 18% of the variance alone.

The results also exhibited some unexpected patterns. For instance, studies into the effects of cohesive devices have suggested that more coherent essays would be produced

Table 6. Test set statistics (*M*, *SD*) for Coh-Metrix indices as a function of grade.

Coh-Metrix index	Assigned grade/rating					
	A	B	C	D	E	F
Lexical diversity *D*	97.238 (22.999)	86.759 (27.775)	83.032 (28.780)	74.103 (22.715)	66.759 (23.556)	64.968 (20.964)
CELEX content word frequency	1.197 (0.197)	1.157 (0.202)	1.275 (0.209)	1.263 (0.234)	1.334 (0.211)	1.336 (0.255)
Word meaningfulness	621.595 (44.477)	640.265 (45.929)	631.731 (47.081)	643.085 (37.863)	641.787 (46.750)	648.768 (40.006)
Aspect repetition	0.909 (0.089)	0.944 (0.073)	0.951 (0.069)	0.949 (0.062)	0.921 (0.090)	0.949 (0.050)
Word familiarity every word	593.736 (2.535)	594.464 (2.927)	594.522 (2.833)	595.548 (2.491)	596.383 (3.449)	596.570 (2.696)

Table 7. Total set statistics (*M*, *SD*) for Coh-Metrix indices as a function of grade.

Coh-Metrix index	Assigned grade/rating					
	A	B	C	D	E	F
Lexical diversity *D*	105.524 (28.854)	84.159 (24.251)	83.043 (23.620)	76.149 (21.894)	69.955 (23.112)	66.527 (20.763)
CELEX content word frequency	1.143 (0.215)	1.172 (0.208)	1.250 (0.206)	1.273 (0.229)	1.331 (0.224)	1.350 (0.237)
Word meaningfulness	617.590 (44.299)	635.727 (42.672)	635.814 (45.533)	639.304 (39.707)	643.904 (44.008)	645.547 (42.035)
Aspect repetition	0.917 (0.081)	0.930 (0.083)	0.947 (0.060)	0.948 (0.062)	0.941 (0.081)	0.952 (0.058)
Average of word familiarity every word	592.880 (2.662)	594.405 (2.742)	594.503 (2.592)	594.962 (2.687)	596.260 (3.012)	596.721 (2.868)

by writers judged to be more proficient (with the exception of studies considering lexical diversity). These findings are premised on the notion that more proficient writers possess the linguistic ability to produce more and varied cohesive devices and better understand the purpose and effects of cohesive devices. However, our study does not support this assertion, with writers judged to be more proficient actually producing texts with fewer cohesive devices. For instance, this study demonstrated that higher-scored essays provided less lexical overlap than lower-proficiency essays. This finding is supported by the *D* (lexical diversity) findings, which demonstrate more lexical diversity at higher proficiency levels as compared with lower proficiency levels. Additional support for this finding can be found in the correlation analysis, which demonstrated that essays written by L2 writers evaluated as less proficient contain more content word overlap and higher semantic similarity scores. Writers judged to be of higher proficiency also provide less aspect repetition than lower-proficiency writers. Aspect helps the reader maintain information in working memory (Magliano & Schleich, 2000) and researchers have argued that more proficient writers would repeat aspect as a method of creating greater cohesion in text (Duran et al., 2007). However, this appears not to be the case with the L2 writers sampled in this study. While not included in the regression analysis, many cohesion variables demonstrated similar trends in the correlation analysis. For example, higher-proficiency writers produce texts with fewer positive logical connectors (e.g. *and*, *also*, *then*, *in sum*, *next*) than lower-proficiency writers. This finding could be counterintuitive because one might expect that more proficient writers would want to make strong links between ideas and clauses (Crismore et al., 1993; Longo, 1994) and provide for a more organised text through the use of connectives (Van de Kopple, 1985). Additionally, writers judged to be of higher proficiency provide readers with less given material than writers judged to be lower proficiency and thus produce texts that require readers to activate more lexical knowledge (Chafe, 1975) and depend less on the preceding discourse (Halliday, 1967). The use of less given information would require more cognitive processing on the part of the reader, something we might expect that higher-proficiency writers would avoid.

Overall, the results of this study suggest that L2 writers judged to be more advanced produce texts with fewer cohesive devices. This finding counters past L2 studies (Connor, 1990; Jin, 2001) that found that more advanced writers used more cohesive devices. However, the use of fewer cohesive devices by writers judged to be more proficient has been supported in L1 writing studies (McNamara et al., 2010). One reason for this might be *a reverse cohesion effect*. Reading comprehension studies (McNamara, Kintsch, Songer & Kintsch, 1996; O'Reilly & McNamara, 2007) have demonstrated that low-knowledge readers benefit more from cohesive texts than high-knowledge readers, who actually benefit more from lower-cohesion texts. Thus, more proficient writers, assuming that their audience includes high-knowledge readers, might produce less cohesive texts.

In contrast to text cohesion, our linguistic sophistication findings adhere to research-supported expectations. For instance, past research has shown that higher-proficiency writers use greater lexical diversity than lower-proficiency writers (see cohesion findings above; Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994). From a lexical difficulty perspective, past research has also demonstrated that higher-proficiency writers also use more infrequent words (Meara & Bell, 2001; Nation, 1988). As in past studies, our findings support the use of greater lexical diversity and less frequent words by L2 writers judged to be more proficient. This study also investigated four additional lexical difficulty measures that have not been considered in the past: word familiarity, word

meaningfulness, word concreteness and word imagability. In reference to word familiarity, higher-proficiency writers use words that are less familiar and thus likely more difficult to recognise. In reference to word meaningfulness, writers evaluated as more proficient use words that are less meaningful and thus have fewer associations with other words, making lexical connections within the text more difficult to develop. Writers judged to be more proficient also produced words that were less concrete and less imageable. In consideration of these findings, we are left with the conclusion that writers judged as more proficient produce more infrequent words that have fewer associations, are less familiar and are more abstract and less imageable. Thus, a mark of writers evaluated as having higher proficiency is an increased level of linguistic sophistication.

The findings regarding linguistic sophistication raise an additional theoretical implication: the relative importance of lexical variables in writing proficiency. Of the five variables in the regression, four of them are lexical (lexical diversity, word frequency, word meaningfulness and word familiarity). These four variables account for almost all of the variance in the regression analysis (29%). Although studies of lexical proficiency have been limited (Engber, 1995; Meara, 2002), the results of previous studies have strongly suggested that lexical knowledge is an important aspect of L2 writing proficiency (Engber, 1995). Additionally, lexical proficiency is a critical factor in the creation of global errors that lead to breaks in L2 communication (De la Fuente, 2002; Ellis, 1995; Ellis, Tanaka & Yamakazi, 1994), especially in timed writing where essay content strongly correlates to lexical output and the production of incorrect lexicon can obscure the meaning of the text and affect the judgements of the grader (Santos, 1988). Timed writing tasks are also important because they better reflect the lexical resources available to L2 learners (Engber, 1995). Other studies (Harley & King, 1989; Linnarud, 1986; McClure, 1991) have also examined correlations between lexical errors and essay scoring and concluded that human judgements of writing proficiency are primarily based on the correct use of the lexicon and the use of a variety of lexical resources. This study provides additional support for the importance of lexical richness and variety in assessing proficient L2 writing. Furthermore, the study reports on additional lexical measures such as word meaningfulness and word familiarity that do not solely examine lexical knowledge based on surface-level features such as lexical diversity and frequency (Polio, 2001).

Conclusion

This study has demonstrated that a combination of surface code, textbase and situation model variables can be used to analyse differences between low- and high-scored essays written by L2 learners. Like first language writers, L2 writers evaluated as being highly proficient do not appear to produce texts that are more cohesive, but instead produce texts that demonstrate more linguistic sophistication (e.g. McNamara et al., 2010). This sophistication can be observed in the production of texts that use less frequent, less familiar and less meaningful words, while also deploying a more diverse range of words. Additionally, writers judged as highly proficient provide readers with less temporal cohesion and word overlap.

The findings of this study support findings from past studies, but this study also presents new data on the use of cohesive devices in L2 writing as well as introduces new indices at the surface code, textbase and situation model level. These findings deserve

additional investigation. Future studies might consider whether similar results occur in other language populations rather than Hong Kong students. Additionally, future studies might consider what features of the analysed texts outside of the linguistic features might play a role in writing proficiency. These features could include error production, contextual factors such as truthfulness and accuracy, world knowledge and rhetorical style. All of these features might help to explain the additional variance not predicted by the linguistic variables examined in this study. Lastly, while this study's focus is mostly on the writer, future studies should consider how linguistic features affect the essay rater through controlled studies that examine the cognitive effects of cohesion and language sophistication on the rater.

Acknowledgements

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Avent, J. & Austermann, S. (2003). Reciprocal scaffolding: A context for communication treatment in aphasia. *Aphasiology*, 17, 397–404.

Baayen, R.H., Piepenbrock, R. & van Rijn, H. (Eds.) (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium.

Bell, J. & Burnaby, B. (1984). *A handbook for ESL literacy*. Toronto: OISE.

Bialystok, E. (1978). A theoretical model of second language learning model. *Language and Learning*, 28, 69–83.

Biesenbach-Lucas, S., Meloni, C. & Weasenforth, D. (2000). Use of cohesive features in ESL students' e-mail and word-processed texts: A comparative study. *Computer Assisted Language Learning*, 13, 221–237.

Brace, N., Kemp, R. & Snelgar, R. (2006). *SPSS for psychologists: A guide to data analysis using SPSS for Windows*. (3rd edn). London: Palgrave.

Brown, D. & Yule, G. (1983). *Teaching the spoken language*. Cambridge: Cambridge University Press.

Carlson, S., Bridgeman, B., Camp, R. & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and non-native speakers of English (TOEFL research rep. no. 19)*. Princeton, NJ: Educational Testing Service.

Chafe, W.L. (1975). Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In: C.N. Li (Ed.), *Subject and topic*. (pp. 26–55). New York: Academic.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd edn). Hillsdale, NJ: Erlbaum.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33, 497–505.

Connor, U. (1984). A study of cohesion and coherence in ESL students' writing. *Papers in Linguistic: International Journal of Human Communication*, 17, 301–316.

Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, 24, 67–87.

Connor, U. (1996). *Contrastive rhetoric*. Cambridge: Cambridge University Press.

Costerman, J. & Fayol, M. (1997). *Processing interclausal relationships: Studies in production and comprehension of text*. Hillsdale, NJ: Lawrence Erlbaum.

Crismore, A., Markkanen, R. & Steffensen, M. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication*, 10, 39–71.

Crossley, S.A., Louwerse, M.M., McCarthy, P.M. & McNamara, D.S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, 91(2), 15–30.

Crossley, S.A. & McNamara, D.S. (2009). Computationally assessing lexical differences in second language writing. *Journal of Second Language Writing*, 17(2), 119–135.

Crossley, S.A., Salsbury, T., McCarthy, P.M. & McNamara, D.S. (2008). Using Latent Semantic Analysis to explore second language lexical development. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 21st International Florida Artificial Intelligence Research Society*. (pp. 136–141). Menlo Park, CA: AAAI Press.

Crossley, S.A, Salsbury, T. & McNamara, D.S. (2009). Measuring second language lexical growth using hypernymic relationships. *Language Learning*, 59(2), 307–334.

Crossley, S.A., Salsbury, T. & McNamara, D.S. (in press). The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3).

Daller, H., van Hout, R. & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24, 197–222.

De la Fuente, M.J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition*, 24, 81–112.

Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.), *Learning to read in different languages*. (pp. 33–102). Washington, DC: Washington Center for Applied Linguistics.

Dufty, D.F., Graesser, A.C., Lightman, E., Crossley, S.A. & McNamara, D.S. (2006). *An algorithm for detecting spatial cohesion in text*. Presentation at the 16th Annual Meeting of the Society for Text and Discourse, Minneapolis, MN.

Dunkelblau, H. (1990). *A contrastive study of the organizational structures and stylistic elements of Chinese and English expository writing by Chinese high school students*. Dissertation Abstracts International, 51(4), 1143A.

Duran, N.D., McCarthy, P.M., Graesser, A.C. & McNamara, D.S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods*, 29, 212–223.

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics*, 16, 409–435.

Ellis, R., Tanaka, Y. & Yamakazi, A. (1994). Classroom interaction, comprehension, and L2 vocabulary acquisition. *Language Learning*, 44, 449–491.

Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139–155.

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

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

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage Publications.

Frase, L., Faletti, J., Ginther, A. & Grant, L. (1997). *Computer analysis of the TOEFL test of written English (TOEFL research rep. no. 64)*. Princeton, NJ: Educational Testing Service.

Gilhooly, K.J. & Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation*, 12, 395–427.

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.

Graesser, A.C., Millis, K.K. & Zwaan, R.A. (1997). Discourse comprehension. *Annual Review of Psychology*, 48, 163–189.

Grant, L. & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9, 123–145.

Grela, B.G. (2002). Lexical diversity in children with Down Syndrome. *Clinical Linguistics and Phonetics*, 16, 251–263.

Halliday, M.A.K. (1967). Notes on transitivity and theme in English. Journal of Linguistics, 3, 199–244.

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

Harley, B. & King, M.L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, 2, 415–440.

Hempelmann, C.F., Dufty, D., McCarthy, P.M., Graesser, A.C., Cai, Z. & McNamara, D.S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In B.G. Bara,

L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. (pp. 941–946). Mahwah, NJ: Erlbaum.

Herskovits, A. (1998). Schematization. In O. Gapp (Ed.), *Presentation and processing of spatial expressions*. (pp. 149–162). Mahwah, NJ: Lawrence Erlbaum Associates.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19, 57–84.

Jarvis, S., Grant, L., Bikowski, D. & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, 12, 377–403.

Jin, W. (2001). *A quantitative study of cohesion in Chinese graduate students' writing: Variations across genres and proficiency levels*. (ERIC document reproduction service no. ED 452 726).

Jurafsky, D. & Martin, J.H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kamel, G. (1989). *Argumentative writing by Arab learners of English as a foreign and second language: An empirical investigation of contrastive rhetoric*. Dissertation Abstracts International, 50(3), 677A.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W. & Van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363–394.

Klein, W. (1994). *Time in language*. London: Routledge.

Kubota, R. (1998). An investigation of L1–L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of Second Language Writing*, 7, 69–100.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211–240.

Landauer, T.K., Foltz, P.W. & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284.

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

Longo, B. (1994). Current research in technical communication: The role of metadiscourse in persuasion. *Technical Communication*, 41, 348–352.

Louwerse, M.M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 12, 291–315.

Louwerse, M.M. (2004). Un modelo conciso de cohesion en el texto y coherencia en la comprehension [A concise model of cohesion in text and coherence in comprehension]. *Revista Signos*, 37, 41–58.

Magliano, J.P. & Schleich, M.C. (2000). Verb aspect and situation models. *Discourse Processes*, 29, 83–112.

Malvern, D.D., Richards, B.J., Chipere, N. & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills: Palgrave Macmillan.

Matsuda, P.K. (1997). Contrastive rhetoric in context: A dynamic model of L2 writing. *Journal of Second Language Writing*, 6(1), 45–60.

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.

McCarthy, P.M. & Jarvis, S. (2007). *vocd:* A theoretical and empirical evaluation. *Language Testing*, 24, 459–488.

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

McCarthy, P.M., Lehenbauer, B.M., Hall, C., Duran, N.D., Fujiwara, Y. & McNamara, D.S. (2007). A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British scientists. *Foreign Languages for Specific Purposes*, 6, 46–77.

McClure, E. (1991). A comparison of lexical strategies in Ll and L2 written English narratives. *Pragmatics and Language Learning*, 2, 141–154.

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

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

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research*, 18(4), 393–407.

Meara, P. & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16(3), 323–337.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990) *Five papers on WordNet.* Cognitive Science Laboratory, Princeton University, no. 43.

Milton, J. (2000). *Research report: Elements of a written interlanguage: A computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students*. Hong Kong: HKUST.

Nation, P. (1988). *Word lists*. Victoria: University of Wellington Press.

Nation, P. & Heatley, A. (1996). *VocabProfile, word and range: Programs for processing text*. LALS, Victoria University, Wellington.

Nunan, D. (1989). *Designing tasks for the classroom*. Cambridge: Cambridge University Press.

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.

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, 4, 32–38.

Pearson, P.D. (1974-75). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relationships. *Reading Research Quarterly*, 10, 155–192.
Perfetti, C.A. (1985). *Reading ability*. Oxford: Oxford University Press.
Perfetti, C.A., Landi, N. & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M.J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook*. (pp. 227–247). Oxford: Blackwell.
Polio, C. (2001). Research methodology in second language writing research: The case text-based studies. In T. Silva & P.K. Matsuda (Eds.), *On second language writing*. (pp. 91–115). Mahwah, NJ: Lawrence Erlbaum Associates.
Rashotte, C.A. & Torgesen, J.K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly*, 20, 180–188.
Rayner, K. & Pollatsek, A. (1994). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.
Reid, J. (1986). Using the writer's workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing*. (pp. 167–188). Alexandria, VA: TESOL.
Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom*. (pp. 191–210). Cambridge: Cambridge University Press.
Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing*, 1, 79–107.
Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective*. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.
Salsbury, T., Crossley, S.A & McNamara, D.S. (in press). Psycholinguistic word information in second language oral discourse. *Second Language Research*, 26(2).
Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, 22, 69–90.
Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27(4), 657–675.
Tabachnick, B.G. & Fidell, L.S. (2001). *Using multivariate statistics*. (4th edn). Needham Heights, MA: Allyn & Bacon.
Templin, M. (1957). *Certain language skills in children*. Minneapolis, MN: University of Minnesota Press.
Toglia, M.P. & Battig, W.R. (1978). *Handbook of semantic word norms*. New York: Erlbaum.
Van de Kopple, W. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication*, 36, 82–95.
White, R. (1981). Approaches to writing. *Guidelines*, 6, 1–11.
Whitten, I.A. & Frank, E. (2005). *Data mining*. San Francisco, CA: Elsevier.
Witte, S.P. & Faigley, L. (1981). Coherence, cohesion and writing quality. *College Composition and Communication*, 22, 189–204.
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 A. Crossley is an Assistant Professor of Applied Linguistics at Georgia State University. His interests include computational linguistics, corpus linguistics, and second language acquisition. He has published articles in second language lexical acquisition, multi-dimensional analysis, discourse processing, speech act classification, cognitive science, and text linguistics.

Danielle S. McNamara is a Professor at the University of Memphis and Director of the Institute for Intelligent Systems. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research ranges a variety of topics including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms.

Received 6 November 2009; revised version received 17 February 2010.

Address for correspondence: Scott A. Crossley, Department of Applied Linguistics & ESL, Georgia State University, P.O. Box 4099 Atlanta, GA 30302-4099, USA. E-mail: *scrossley@lgsu.edu*