Computational assessment of lexical differences in L1 and L2 writing

Scott A. Crossley a,*, Danielle S. McNamara b,1

a Department of English, Mississippi State University, P.O. Box E., Mississippi State University, MS 39762, United States
b Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, United States

Abstract

The purpose of this paper is to provide a detailed analysis of how lexical differences related to cohesion and connectionist models can distinguish first language (L1) writers of English from second language (L2) writers of English. Key to this analysis is the use of the computational tool Coh-Metrix, which measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis, and a statistical method known as discriminant function analysis. Results show that L1 and L2 written texts vary in several dimensions related to the writer’s use of lexical choices. These dimensions correlate to lexical depth of knowledge, variation, and sophistication. These findings, together with the relevance of the new computational tools for the text analysis used in the study, are discussed.

© 2009 Elsevier Inc. All rights reserved.

Keywords: Second language writing; Lexical proficiency; Corpus linguistics; Computational linguistics; Cohesion; Lexical networks

Introduction

The purpose of this study is to explore how lexical differences related to cohesion and lexical networks can be used to distinguish between texts written by first language (L1) writers of English and second language (L2) writers of English. As such, the study is relevant because it compares L1 and L2 writing samples based on lexical features, an area of study that has been traditionally overlooked (Meara, 2002). Additionally, this study provides an overview of a new computational tool, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) that allows for the deep analysis of large corpora on several dimensions of language, discourse, and conceptual analysis. The tools and the statistical methods used in this study are important not only because they distinguish how L2 writers differ in their use of lexical features, but also because past research has concluded that approaches relying on lexical variables alone could not be used to discriminate L1 and L2 texts (Connor, 1984; Reynolds, 1995). Our research, however, demonstrates that using computational tools to analyze lexical features results in statistically significant findings that provide strong evidence of the different lexical knowledge available to L1 and L2 writers. This evidence supports notions of lexical differences between L1 and L2 populations, but, more importantly, it describes them in reference to their cohesive effects and their networking properties.

In what follows we present an overview of the relevant empirical literature on the dimension of writing covered in our study, and of the computational tool used in the analysis of our data. We provide this overview in order to conceptualize our own research and the research questions guiding it.

* Corresponding author. Tel.: +1 662 325 2355.
E-mail addresses: sc544@msstate.edu (S.A. Crossley), d.mcnamara@mail.psyc.memphis.edu (D.S. McNamara).
1 Tel.: +1 901 678 3803.

1060-3743/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
Differences between L1 and L2 writing

Distinguishing between texts written by L1 and L2 writers has a long history in second language writing studies. In an overview of research concerning differences between English L1 and L2 writing, Silva (1993) found a number of salient differences. These differences include the notion that L2 writers use more but shorter T units (main clauses and dependent clauses), fewer but longer clauses, more coordination, less subordination, less noun modification, less passivization, and fewer cohesion devices including more conjunctives and lexical ties. Concerning the L2 writers’ use of lexicon specifically, Silva’s overview found that L2 writers demonstrated less lexical control, variety, and overall lexical sophistication. Individual studies have also provided insight into the lexical differences between L1 and L2 writers and speakers. For instance, Spanish writers of English have been shown to use longer sentences, more pronouns (Reid, 1988), more nouns, more lexical overgeneralizations (McClure, 1991), fewer simple sentences, more synonyms, and more additive and casual conjunctions (Connor, 1984). Swedish writers of English have shown to differ from English writers in text length, error productions, lack of variation, originality in vocabulary, use of idioms and collocations, and word frequency. These findings have been supported in other studies (e.g., Harley & King, 1989; McClure, 1991). Taken together, the general conclusion from these studies is that texts produced by L2 writers generally have less lexical variety, specificity, and sophistication than those written by L1 writers.

Lexical proficiency

Historically the study of lexicon and second language acquisition has been neglected in second language research (Meara, 2002) and, as a result, examining texts based on lexical features is a less common approach than measuring linguistic accuracy or complexity. Of the studies that have examined lexical proficiency, much of the research has considered surface measures of lexical proficiency such as lexical originality, density, accuracy, and diversity (Polio, 2001). However, these surface measures do not provide insight into deeper, cognitive measures of L2 lexicons such as the development of word senses and lexical networks (Crossley, Salsbury, & McNamara, 2009; Schmitt, 1998). The examination of lexical features related to word sense development and lexical networks is important because words are more than just simple, linear connections between form and meaning. Thus L2 lexical proficiency cannot be evaluated solely on basic performance rubrics such as the number of words used or comprehended or the ability to match dictionary definitions with strings of letters (Crossley et al., 2009). To truly understand an L2 learner’s lexical proficiency, we must move past simple analyses of lexical features and begin to examine the L2 learner’s lexical knowledge of syntagmatic and paradigmatic properties, sense relations, and complex lexical association models (Haastrup & Henriksen, 2000; Huckin & Coady, 1999).

Understanding learner’s lexical proficiency is crucial because L2 lexical errors often produce global errors that inhibit communication (de la Fuente, 2002; Ellis, 1995; Ellis, Tanaka, & Yamakazi, 1994). In addition, lexical growth strongly correlates with academic achievement (Daller, van Hout, & Treffers-Daller, 2003). Thus, studies that look at lexical production at a holistic level can lead to a better appreciation of the lexicon as a primary feature of language development. Understanding lexical acquisition in relation to its deeper, cognitive functions is also important because it can lead to increased awareness of how learners process and produce a second language. Currently such studies are rare and an overall theory of lexical acquisition is absent (Nation, 1990). We attempt to address this research gap by analyzing the lexical differences between L1 and L2 writers at two cognitive levels: cohesion and lexical networks with the working theory that differences between the two will highlight the restricted lexical proficiency common in L2 learners.

Cohesion

Textual cohesion is a critical aspect of successful language processing and comprehension and is premised on building connections between ideas in text. As noted by both Silva (1993) and Ferris (1994), cohesion plays an important part in the lexical development of L2 writers, and it also serves as a means to distinguish differences between L1 and L2 writers. Historically, there has been some debate about identifying the relationship between cohesion and coherence. This paper will adapt the psycholinguistic distinction put forth by Louwerse (2004) in which coherence refers to the representational relationships of a text in the mind of a reader whereas cohesion refers to the textual indications that coherent texts are built upon. In essence, then, cohesion consists of the elements of the text, while
coherence refers to the consistency of the elements as a mental representation. The more cohesive devices in a text, the more coherent it will be and the easier it will be to understand.

Early research examining the use of surface-level cohesive devices in L2 writing has demonstrated that L2 writers differ from L1 writers in important ways. Connor (1984), for example, found that the density of overall cohesive devices was not enough to differentiate L1 and L2 essays. However, using just one type of cohesive device, lexical cohesion, did distinguish L1 and L2 writing because L2 writers had a higher percentage of lexical reiteration and fewer collocations and synonyms. Reid (1992) looked at cohesive devices in a corpus of 768 essays written in English by L1 and L2 writers and analyzed them for personal and demonstrative pronouns (referential cohesion) and the use of coordinate and subordinate conjunctions (conjunctive devices). Reid found significant differences between L1 writers and various L2 groups. Specifically, native speakers used far fewer pronouns and coordinating conjunctions than the L2 groups, and far more subordinate conjunctions. Similar findings have been reported by Scarcella (1984), Johns (1984), and Ventola and Muraanen (1991). These studies, taken together, support the notion that the effectiveness of L2 written texts may be impaired by the writer’s use of cohesive devices.

Lexical networks

Lexical acquisition is now recognized to be a complex phenomenon that goes well beyond the memorization of words and definitions (Bogaards, 2001; Crossley, Salsbury, McCarthy, & McNamara, 2008b; Crossley et al., 2009). Word knowledge and word acquisition involve the creation of connections between related words (Haastrup & Henriksen, 2000; Huckin & Coady, 1999). The linking of words based on similar concepts, senses, and relations is based on the notion of lexical networks and predicated on the interrelation between words to form categorical clusters. Once clusters are formed, they begin to connect with other clusters, creating lexicons based on shared connections (Ferrer i Cancho & Solé, 2001; Ferrer i Cancho, Solé, & Köhler, 2004; Haastrup & Henriksen, 2000). The value of these connections is that they allow for relatively quick growth of lexicons as new words can be assimilated with known words. As more words are associated with each other, the lexical network strengthens (Haastrup & Henriksen, 2000).

Since theories of lexical networks are relatively recent, little research comparing differences between L1 and L2 lexical networks exists. However, a few recent articles have shown that L2 learners’ lexical networks contain more random associations than L1 networks (Meara & Schur, 2002) and that L1 speakers provide more word associations in lexical knowledge tests than do L2 learners (Zareva, 2007). Other recent studies have demonstrated that lexical networks develop quite rapidly in L2 learners and that these lexical networks allow for greater lexical diversity, abstractness (Crossley et al., 2009), and lexical coreferentiality (Crossley et al., 2008b). All of these factors are important for language processing.

Computational approaches to L2 writing

Because the approach for distinguishing L1 and L2 texts in this study introduces a new computational tool, it is important to discuss how computational tools can aid in the study of L2 writing patterns. While not the norm, computational tools are helpful because many past discourse studies in contrastive rhetoric have either depended on hand counts of linguistic features and rhetorical structures or intuitive comparisons between texts (Reid, 1992). While this is a reputable approach to investigating writing patterns, the results of the analyses may be hindered by the amount of text that can be examined, the fallibility of hand counts, and the subjective nature of intuitive judgments. Computational tools are not without fault, though, and many researchers have reported problems with computer analyses (e.g., Ferris, 1993; Frase, Faletti, Ginther, & Grant, 1997; Granger, 2002). However, as computational tools expand and develop, they begin to provide a more accessible and theoretically sound approach for the quantitative evaluation of texts.

Perhaps the earliest attempt at using a computational tool was Reid’s (1992) study, in which the WWB program STYLEFILES was used. Reid measured four types of linguistic variables in a corpus of 768 essays written in English by L1 writers and L2 writers. This computational approach allowed her to categorize differences between L1 and L2

---

2 This definition differs from the one given by Connor (1996) in which she states that cohesion is based on lexical and grammatical intersentential relationships that are overt, while coherence is based on semantic relationships.
writers based on surface level cohesion features in the text. In a later study, Grant and Ginther (2000) used a computerized tagging system to analyze the proficiency levels of different L2 learners’ essays and found that by using lexical proficiency (type/token ratio and word length), lexical features (e.g., conjuncts, hedges, amplifiers, and emphatics), grammatical features (e.g., nouns, verbs, nominalizations, and modals), and clause level features (e.g., subordinations, complementation, and passives), they were able to obtain detailed information about proficiency levels. Specifically, as the proficiency level of the learner increased, so did the type/token ratio, average word length, conjuncts, amplifiers, emphatics, demonstratives, and downtoners. This led Grant and Ginther to conclude that as writers’ proficiency levels increased, so did the use of lexical features.

Coh-Metrix

Of interest to this study is the use of the computational tool Coh-Metrix, which is capable of measuring cohesion and text difficulty at various levels of language, discourse, and conceptual analysis. As reported in Graesser et al. (2004), recent advances in various disciplines have made it possible to computationally investigate various measures of text and language comprehension that supersede surface components of language and instead explore deeper, more global attributes of language. The various disciplines and approaches that have made this approach possible include computational linguistics, corpus linguistics, information extraction, information retrieval, psycholinguistics, and discourse processing. Taken together, the improvements in these fields allow for the analysis of many deep level factors of textual coherence and language to be automated, providing more accurate and detailed analyses of language.

A synthesis of the developments in these areas has been achieved in Coh-Metrix. The system integrates lexicons, pattern classifiers, part-of-speech taggers, shallow semantic interpreters, and other components that have been developed in the field of computational linguistics (Jurafsky & Martin, 2002). Utilizing these resources, Coh-Metrix can analyze texts on several dimensions of cohesion including co-referential cohesion, causal cohesion, density of connectives, latent semantic analysis metrics, and syntactic complexity. It also includes several lexical metrics such as word frequency, concreteness, polysemy, word meaningfulness, hypernymy, word age-of-acquisition scores, word imagability, and word familiarity measures (Graesser et al., 2004).

Coh-Metrix has been used to distinguish text types, examine the linguistic structures of texts, and explore textual differences in L2 discourse studies (Crossley, Louwerse, McCarthy, & McNamara, 2007; Crossley, McCarthy, & McNamara, 2007; Crossley & McNamara, 2008; McCarthy, Lehenbauer, Hall, Duran, Fujiwara, & McNamara, 2007) and L1 discourse studies (Louwerse, McCarthy, McNamara, & Graesser, 2004; McCarthy, Lewis, Dufty, & McNamara, 2006). In addition, multiple validation studies have been conducted on Coh-Metrix and Coh-Metrix measures in relation to cohesion (e.g., Crossley, Greenfield, & McNamara, 2008; Crossley et al., 2008b; Crossley, Salsbury, McCarthy, & McNamara, 2008a; Dufty, Graesser, Lightman, Crossley, & McNamara, 2006; Dufty, Graesser, Louwerse, & McNamara, 2006; Dufty, Hempelmann, Graesser, Cai, & McNamara, 2005; Hempelmann, Dufty, McCarthy, Graesser, Cai, & McNamara, 2005; McNamara, Ozuru, Graesser, & Louwerse, 2006) and lexical indices (Crossley et al., 2008b, 2009).

In the study to be reported next, the hypothesis that lexical differences exist between L1 writers and L2 writers of English was empirically tested. In order to do so, we collected a corpus of L1 and L2 texts and used the findings from Coh-Metrix as a foundation for a statistical analysis. More precisely, using corpus, computational, and statistical approaches, this study examined whether lexical difference can be used to distinguish L1 essays from L2 essays and what useful information about L2 lexical proficiency can be gained from these findings.

Method

Corpus selection

Two corpora were used for this study: a corpus of essays written in English by native Spanish speakers taken from the International Corpus of Learner English (ICLE) and a matching L1 corpus collected from undergraduate students at Mississippi State University. The ICLE was designed with strict criteria including learner level and rhetorical style. These criteria were implemented in order to make data interpretation easier and allow for clear conclusions as to the kind of errors or differences produced and under what conditions (Granger, Dagneaux, & Meunier, 2002). The ICLE was designed to consider learner variables such as age (university students in their twenties), learning context (EFL),
proficiency (high intermediate to advanced writers), and mother tongue. It was also designed to consider task variables such as medium (writing), genre (academic essays), field (general), and essay length (between 500 and 1000 words). The majority of the essays contained in the ICLE are argumentative essays that allow for discourse-orientated as well as grammatical and lexical investigation. This is important for this study because lexical cohesive devices (discourse-oriented variables) as well as other lexical indices are examined.

The corpus collected at Mississippi State University was designed to closely follow the criteria used for the ICLE corpus, except for the second language criterion. The 208 essays that comprised the L1 corpus were collected from native English speaking undergraduate college students in persuasive writing classes. The essays collected were all argumentative essays that followed the four most common essay topics found in the ICLE corpus (see Table 1). As a result of instruction to the writers, all the essays ranged between 500 and 1000 words. The essays were purposefully kept to a similar length as the ICLE essays given that longer essays could have greater coverage of lexical items and might allow for lexical items to be repeated (Engber, 1995), possibly skewing the data. Like the majority of the ICLE essays, the essays were also untimed essays written outside of the classroom. This means that referencing of outside sources was allowed, but was not necessary.

Not all the available essays were taken from the Spanish corpus of the ICLE. Of the available 251 essays, only those that were argumentative essays were selected. Because of formatting differences, only 195 of the available 199 essays were successfully processed through Coh-Metrix. The discourse features of the 208 L1 texts were used as the norm and compared to the 195 Spanish texts in the ICLE corpus. This was done to determine in what ways L2 writers of English differ from L1 writers in their use of lexical cohesive devices and other lexical features. This is a similar methodology to that used by Reid (1992) and like Reid, the use of the L1 texts should be seen as a model of text, not an ideal text. An overview of the two corpora can be found in Table 2. It is worth mentioning that while the differences between the text length in the L1 and L2 corpora appear relatively large, the measures Coh-Metrix reports are normalized for length.

Data analyses

To test the hypothesis that there are linguistic differences that differentiate L1 texts from L2 texts, we conducted a discriminant function analysis. Considering the work that had been conducted in the past in measuring the lexical ability of L2 writers, especially in distinguishing between L1 and L2 writers, only those Coh-Metrix indices that measure lexical features related to cohesion and lexical networks were selected. All the selected indices had been validated in past studies. The indices selected for this study were lexical coreferentiality, semantic coreferentiality (latent semantic analysis), word frequency measures, word information measures from the MRC psycholinguistic

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Most common essay topics in ICLE.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?</td>
<td>452</td>
</tr>
<tr>
<td>Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television.</td>
<td>231</td>
</tr>
<tr>
<td>In his novel ‘Animal Farm’, George Orwell wrote “All men are equal: but some are more equal than others”. How true is this today?</td>
<td>115</td>
</tr>
<tr>
<td>Feminists have done more harm to the cause of women than good.</td>
<td>105</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Corpus information.</th>
</tr>
</thead>
<tbody>
<tr>
<td>English essays</td>
<td>Spanish essays</td>
</tr>
<tr>
<td>Number of essays</td>
<td>208</td>
</tr>
<tr>
<td>Mean number of words</td>
<td>726.183</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>128.759</td>
</tr>
<tr>
<td>Total words in corpus</td>
<td>151,046</td>
</tr>
</tbody>
</table>

database, hypernymy and polysemy values, spatiality, and causality. These indices are briefly discussed below. More detailed information can be found in Graesser et al. (2004).

**Coreferentiality**

Coh-Metrix considers three forms of lexical co-reference between sentences: noun overlap between sentences, argument overlap between sentences, and stem overlap between sentences. Noun overlap measures how often a common noun is shared between two sentences. Argument overlap measures how often two sentences share nouns with common stems, while stem overlap measures how often a noun in one sentence shares a common stem with other word types in another sentence. When links are made between concepts at the suprasentential level, larger units of meaning are constructed. Thus the meaning of text does not reside solely in the sum meaning of the individual sentences, but in how those sentences are connected together (Just & Carpenter, 1987; Rayner & Pollatsek, 1994). These types of cohesive links have been shown to aid in text comprehension and reading speed (Douglas, 1981; Kintsch & Van Dijk, 1978; Rashotte & Torgesen, 1985). Lexical coreferentiality indices are also of interest in L2 writing because L2 writers have been shown to use a higher percentage of lexical repetition (Connor, 1984) and advanced L2 writers have been found to use a variety of lexical and referential cohesion devices, while lower level writers have been reported to use more repetition (Ferris, 1994). The inclusion of coreferentiality variables is also of interest in light of Reynolds’ (1995) study, which attempted to compare the amount of repetition between L1 and L2 written texts using objective criteria. Reynolds reported that the tracking of either the underuse or overuse of repetition had not been historically possible, and even his 1995 study demonstrated that distinguishing between L1 and L2 writers using repetition was not possible, leaving him to contend that such efforts were unproductive.

**Latent semantic analysis (LSA)**

Coh-Metrix tracks semantic coreferentiality using LSA, which is a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts. LSA uses a general form of factor analysis to condense a very large corpus of texts 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 ends up being a weighted vector (Landauer and Dumais, 1997; Landauer, Foltz, & Laham, 1998). Unlike lexical markers of coreferentiality, LSA provides for the tracking of words that are semantically similar, but may not be related morphologically. For instance, the word *mouse* has a higher LSA score when compared to *cat* than to either *dog* or *house*. In addition, Coh-Metrix tracks givenness through LSA by measuring the proportion of new information each sentence provides according to LSA. Given information is thought to be recoverable from the preceding discourse (Halliday, 1967) and does not require activation (Chafe, 1975). Given information is thus less taxing on the reader’s cognitive load. To compute the LSA givenness 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 et al., 2005). LSA is an important measure of L2 cohesion because it can help measure the amount of semantic coreferentiality in a text (Crossley, Louwerse, et al., 2007; Crossley, McCarthy, et al., 2007). LSA is also related to the development of lexical networks in L2 learners (Crossley et al., 2008b).

**Word frequency**

Word frequency in Coh-Metrix refers to metrics of how frequently particular words occur in the English language. For instance, the word *mother* would be more frequent than the word *cousin*. 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. The tracking of word frequency is important because L2 writers are assumed to use less lexical variety and sophistication (Linnarud, 1986; Silva, 1993). From a cohesion perspective, more frequent words allow for quicker decoding (Perfetti, 1985; Rayner & Pollatsek, 1994). When decoding is automatic, it places smaller demands on a reader’s working memory. If decoding is not automatic, then working memory processes are dedicated to decoding and not to comprehension. This can affect the retention of textual information (Field, 2004). Frequent words are also processed more quickly and understood better than infrequent words (Haberlandt & Graesser, 1985; Just & Carpenter, 1980).
Word information (MRC psycholinguistic database)

Coh-Metrix calculates word information on five psycholinguistic, lexical 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 psychological properties. An example of a high imagery word is buffalo as compared to the low imagery score for relevant. An example of meaningful word is person as compared to amorphous. Words with higher meaningfulness scores activate more related words than those with low meaningfulness scores. A high age-of-acquisition score can be found for a word like alacrity, but not for afternoon. Words with higher age-of-acquisition scores denote spoken words that are learned later by L1 children. A word scoring high on familiarity would be eat, while adze would score low. Familiar words are words that are processed more quickly. Words that score low on the concreteness scale include admire, dignity, and honesty. Portrait, restaurant, and shrimp score high on the concreteness scale. Word that are more concrete are those things you can hear, smell, touch, taste and feel and are processed more quickly.

The MRC database contains 150,837 words and provides information for up to 26 different linguistic properties of these words. Unlike other electronic dictionaries, the MRC psycholinguistic database does not attempt to provide semantic information about the words it contains and, because most MRC measures are based on psycholinguistic experiments, the coverage of words differs among the measures (e.g., the database contains 4825 words with imagery ratings and 4920 with familiarity ratings). Many of these indices are important for L2 lexical networks and lexical cohesion because they can measure how words prime other words, how familiar words are, and how highly words associate with other words.

Hypernymy and polysemy indices

Coh-Metrix tracks the relative ambiguity of a text by calculating its lexical polysemy value, which refers to the number of meanings or senses within a word. Coh-Metrix tracks the relative abstractness of a text by calculating its lexical hypernymy value, which refers to the number of levels a word has in a conceptual, taxonomic hierarchy. 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 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 grape, which has only one sense. Both hypernymy and polysemy values also relate to the development of L2 lexical networks. Hypernymy values can track the growth of lexical connections between hierarchically related items (vehicle and car), while polysemy values can measure the development of sense relations (Crossley et al., 2009).

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 et al., 2006; Dufty, Graesser, Louwerse et al., 2006). Classifications for both motion verbs and location nouns can be found in WordNet (Fellbaum, 1998). Spatial cohesion is predicated on recent theories of discourse processing, which assume that discourse is represented in a situation model by a small number of dimensions including spatial information. Spatial cohesion helps to construct a text and ensures that the situational model of the text (Kintsch & Van Dijk, 1978) is well structured and clearly conveys the text meaning to the reader.

Causal cohesion

Along with spatiality, causality helps to construct the situation model of a text (Kintsch & Van Dijk, 1978). Causal cohesion is measured in Coh-Metrix by calculating the ratio of causal verbs to causal particles (Dufty et al., 2005). 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, consequence of, and as a result. Causality 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 in order to show causal relationship between simple clauses (Pearson, 1974–1975). This variable is of interest in consideration of Reynolds’ (2002) study, which also looked at causality markers in L1.
and L2 essays. In his study, Reynolds’ noted no transition from oral to written usage of causal connectives. However, differences were noted in the use of these connectives between L1 and L2 writers. Reynolds speculated that this difference might be the result of L2 writers depending on narrative prose.

Variable selection

To select the variables from the chosen Coh-Metrix indices, we randomly divided the L1 and L2 essays into approximately equal groups, hereafter the training set (n = 201 texts) and the test set (n = 202 texts). The purpose of the training set was to identify which of the variables contained within the chosen Coh-Metrix indices (lexical coreferentiality, semantic coreferentiality, word frequency measures, word information measures from the MRC psycholinguistic database, hypernymy and polysemy values, spatiality, and causality) best distinguished the L1 from the L2 essays. These selected variables were later used to predict the L1 from the L2 essays in the training set using a discriminant function analysis (detailed below). The L1 and L2 essays in the test set data were later assessed by the discriminant function analysis model.

A discriminant function analysis is a common approach used in many previous studies that distinguished text types (e.g., Biber, 1993; Crossley, McCarthy, & McNamara, 2007). Considering that the training set and the test set contained at least 200 writing samples each and using a conservative estimate of one predictor per 20 variables, we determined that 10 indices would be an appropriate number of predictors for the discriminant analysis that would not create problems of overfitting. Such a ratio is standard for analyses of this kind (Field, 2004). Protecting a model against over-fitting is important because if too many variables are used the model fits not just the signal of the predictors but also the unwanted noise. When overfitting is a problem, the training model fits the data well, but when this model is applied to new data the fit lacks accuracy because the noise will not be the same from data set to data set.

An ANOVA was conducted using the selected eight Coh-Metrix indices as the dependent variables and the L1 and L2 essays from the training set as the independent variables. Tests for homogeneity of variance were also conducted to ensure that the variance between the L1 variables and L2 variables was equal. If the variance was not equal, a likely scenario considering the large sample size and the assumed instability of L2 interlanguages, a Welch’s F score was calculated. We selected the variable with the largest effect size as the representative variable for that index. To obtain the other two variables (for the total of 10), we ranked the remaining variables based on their effect sizes. Not wanting to run the risk of collinearity between variables, which would waste potential model power, we did not simply select the variables with the remaining highest effect sizes. Instead, we tested the variables for collinearity to ensure that no index pair correlated above $r > .70$. If variables that correlated above .70 were used in the model, it would make interpretation difficult because it would be unclear which variables were contributing to the model, as many of the variables might be redundant (Brace, Kemp, & Snelgar, 2006; Tabachnick & Fidell, 2001).

Results

ANOVA

We conducted an ANOVA to select the 10 lexical variables from Coh-Metrix with the highest effect sizes. These variables best distinguished L1 from L2 written essays and were used to demonstrate differences between L1 and L2 essays. The selected variables are discussed below in relation to the information they provide about the differences between L1 and L2 texts (see Table 3 and the section below for additional information).

Word hypernymy

The results of the hypernymy scores reveal that L1 writers use significantly more words that are conceptually more abstract and hierarchically connected than L2 writers of English. This suggests that L1 writers are not as concrete in their writing as L2 writers. It also suggests that L1 writers use greater lexical connections between words.

---

3 The causality markers Reynolds chose came from Kress’ (1994) study in which two types of causal connectors were posited: spatial/temporal (“and,” “then,” and “when”) and agent (“because,” “so,” “therefore,” and “thus”). The first group is frequent in face-to-face conversation, while the second is common in written genres.
The results demonstrate that L1 writers use significantly more words with multiple senses than L2 writers of English. This suggests that L1 writers use words with more senses than L2 writers. It also supports the notion that L1 writers have more lexical associations within core words.

**Argument overlap**

The results show that L1 writers use significantly more argument overlap than L2 writers of English. This suggests that L1 writers provide readers with more lexical coreferentiality than L2 writers and thus produce texts that are more cohesive.

**CELEX written frequency**

The results show that L2 writers are significantly more likely to create texts using more frequent written words than L1 writers. This suggests that the words used by L2 writers are easier to process.

**Age of acquisition**

The findings reveal that L2 writers use higher age-of-acquisition words than L1 writers. This finding suggests that L2 writers are more likely to use words that are considered to be acquired later by children.

**Locational nouns**

The findings demonstrate that L2 writers use significantly more locational nouns than L1 writers. This suggests that L2 writers depend more on locational nouns to provide a sense of spatiality than do L1 writers.

**Word meaningfulness**

This finding demonstrates that L1 writers use words that are significantly more “meaningful” than L2 writers of English. This suggests that L1 writers use more words that activate the recall of other words through closer word associations.

**Incidence of causal verbs**

This finding reveals that L1 writers use significantly more causal verbs than L2 writers. This suggests that English writers provide more cause-and-effect relationships in their writings and thus provide writings that are more cohesive.

---

**Table 3**

<table>
<thead>
<tr>
<th></th>
<th>Argument overlap</th>
<th>Motion verbs</th>
<th>CELEX frequency</th>
<th>LSA givenness</th>
<th>Causal verbs</th>
<th>Age of acquisition</th>
<th>Word meaningfulness</th>
<th>Locational nouns</th>
<th>Polysemy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernymy</td>
<td>0.40</td>
<td>0.02</td>
<td>-0.22</td>
<td>0.34</td>
<td>0.14</td>
<td>-0.23</td>
<td>0.37</td>
<td>-0.33</td>
<td>0.11</td>
</tr>
<tr>
<td>Argument overlap</td>
<td>0.02</td>
<td>0.07</td>
<td>0.39</td>
<td>-0.08</td>
<td>-0.23</td>
<td>0.20</td>
<td>-0.14</td>
<td>-0.09</td>
<td></td>
</tr>
<tr>
<td>Motion verbs</td>
<td>-0.44</td>
<td>-0.02</td>
<td>0.22</td>
<td>0.08</td>
<td>0.13</td>
<td>-0.23</td>
<td>0.36</td>
<td>-0.25</td>
<td></td>
</tr>
<tr>
<td>CELEX frequency</td>
<td>-0.04</td>
<td>-0.23</td>
<td>-0.19</td>
<td>-0.21</td>
<td>-0.17</td>
<td>-0.17</td>
<td>0.55</td>
<td>-0.33</td>
<td></td>
</tr>
<tr>
<td>LSA givenness</td>
<td>0.41</td>
<td>-0.05</td>
<td>0.17</td>
<td>-0.32</td>
<td>0.00</td>
<td>-0.12</td>
<td>-0.01</td>
<td>-0.10</td>
<td></td>
</tr>
<tr>
<td>Causal verbs</td>
<td>-0.03</td>
<td>0.15</td>
<td>0.19</td>
<td>0.23</td>
<td>0.23</td>
<td>-0.33</td>
<td>0.55</td>
<td>-0.24</td>
<td></td>
</tr>
<tr>
<td>Age of acquisition</td>
<td>-0.32</td>
<td>-0.23</td>
<td>-0.19</td>
<td>-0.32</td>
<td>0.00</td>
<td>-0.17</td>
<td>-0.17</td>
<td>-0.10</td>
<td></td>
</tr>
<tr>
<td>Word meaningfulness</td>
<td>-0.01</td>
<td>-0.12</td>
<td>0.17</td>
<td>0.23</td>
<td>0.23</td>
<td>0.55</td>
<td>0.21</td>
<td>-0.33</td>
<td></td>
</tr>
<tr>
<td>Locational nouns</td>
<td>-0.10</td>
<td>-0.32</td>
<td>-0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.55</td>
<td>0.21</td>
<td>-0.33</td>
<td></td>
</tr>
</tbody>
</table>

---

**Note:** The table above includes correlations between selected variables. The values indicate the degree of correlation between each pair of variables, with higher absolute values indicating stronger correlations. The table provides insights into how different features of written texts correlate with each other, helping to identify patterns in L1 and L2 writing styles.
Correlations

None of the variables used in this analysis were highly correlated (see Table 4 for correlations). Thus, we have confidence that no redundant variables were used in this analysis.

Accuracy of model

To test the accuracy of these lexical indices to distinguish between L1 and L2 essays, we conducted a series of discriminant function analyses. A discriminant analysis is a statistical procedure that is able to predict group membership (in this case L1 and L2 essays) using a series of independent variables (in this case the selected Coh-Metrix variables). The training set is used to generate a discriminant function. The discriminant function acts as the algorithm that predicts group membership. This is done first with the training set. Later, the discriminant function analysis model from the training set is used to predict group membership of the essays in the test set. If the results of the discriminant analysis are statistically significant, then the findings support the predictions of the analysis (that linguistic differences exist between L1 and L2 texts and that those differences can be used to classify the texts based on language differences).

In this study, a discriminant function analysis was first conducted using the 10 selected variables as independent variables and the essays (L1 or L2) as the dependent variables. While this allowed for a 20:1 ratio between predictors and items, we did not assume that this was the best possible ratio. Thus, we ran a series of discriminant function analyses to determine if fewer variables would be more predictive. Our first discriminant function analysis, therefore, used 10 variables. We then conducted nine additional discriminant analyses using nine variables through one variable. In each of these discriminant function analyses the variables with the highest effect sizes were kept. The overall accuracy of each discriminant functional analysis can be found in Table 5. The discriminant function analysis that used seven variables (hyponymy, argument overlap, motion verbs, word frequency, polysemy, LSA givenness, and age-of-acquisition scores)

<table>
<thead>
<tr>
<th>Variables</th>
<th>English essays</th>
<th>Spanish essays</th>
<th>( F(1, 112) )</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word hyponymy</td>
<td>5.124 (1.456)</td>
<td>4.001 (1.201)</td>
<td>34.907(^a)</td>
<td>0.149</td>
</tr>
<tr>
<td>Word polysemy</td>
<td>12.900 (2.218)</td>
<td>11.461 (3.437)</td>
<td>34.148</td>
<td>0.149</td>
</tr>
<tr>
<td>Argument overlap</td>
<td>0.170 (.058)</td>
<td>0.134 (.047)</td>
<td>19.807</td>
<td>0.103</td>
</tr>
<tr>
<td>Number of motion verbs</td>
<td>123.175 (55.420)</td>
<td>89.794 (0.764)</td>
<td>18.398</td>
<td>0.086</td>
</tr>
<tr>
<td>CELEX written frequency</td>
<td>3.155 (.081)</td>
<td>2.521 (.076)</td>
<td>13.439</td>
<td>0.064</td>
</tr>
<tr>
<td>Age of acquisition</td>
<td>356.096 (22.076)</td>
<td>369.382 (25.866)</td>
<td>13.163</td>
<td>0.063</td>
</tr>
<tr>
<td>Locational Nouns</td>
<td>219.132 (62.327)</td>
<td>251.905 (63.124)</td>
<td>10.042(^a)</td>
<td>0.049</td>
</tr>
<tr>
<td>LSA givenness</td>
<td>0.345 (.032)</td>
<td>0.325 (.045)</td>
<td>9.384</td>
<td>0.046</td>
</tr>
<tr>
<td>Colorado meaningfulness</td>
<td>429.360 (11.141)</td>
<td>423.438 (14.844)</td>
<td>6.768</td>
<td>0.029</td>
</tr>
<tr>
<td>Incidence of causal verbs</td>
<td>12.375 (6.964)</td>
<td>8.910 (8.857)</td>
<td>4.123</td>
<td>0.016</td>
</tr>
</tbody>
</table>

\(^a\) Welch’s \( F \) value.

Note. All \( F \) values are significant at \( p < .001 \).

Table 4
Means (standard deviations), \( F \) values, and effect sizes for English and Spanish essays.

<table>
<thead>
<tr>
<th>Variables</th>
<th>English essays</th>
<th>Spanish essays</th>
<th>( F(1, 112) )</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word hyponymy</td>
<td>5.124 (1.456)</td>
<td>4.001 (1.201)</td>
<td>34.907(^a)</td>
<td>0.149</td>
</tr>
<tr>
<td>Word polysemy</td>
<td>12.900 (2.218)</td>
<td>11.461 (3.437)</td>
<td>34.148</td>
<td>0.149</td>
</tr>
<tr>
<td>Argument overlap</td>
<td>0.170 (.058)</td>
<td>0.134 (.047)</td>
<td>19.807</td>
<td>0.103</td>
</tr>
<tr>
<td>Number of motion verbs</td>
<td>123.175 (55.420)</td>
<td>89.794 (0.764)</td>
<td>18.398</td>
<td>0.086</td>
</tr>
<tr>
<td>CELEX written frequency</td>
<td>3.155 (.081)</td>
<td>2.521 (.076)</td>
<td>13.439</td>
<td>0.064</td>
</tr>
<tr>
<td>Age of acquisition</td>
<td>356.096 (22.076)</td>
<td>369.382 (25.866)</td>
<td>13.163</td>
<td>0.063</td>
</tr>
<tr>
<td>Locational Nouns</td>
<td>219.132 (62.327)</td>
<td>251.905 (63.124)</td>
<td>10.042(^a)</td>
<td>0.049</td>
</tr>
<tr>
<td>LSA givenness</td>
<td>0.345 (.032)</td>
<td>0.325 (.045)</td>
<td>9.384</td>
<td>0.046</td>
</tr>
<tr>
<td>Colorado meaningfulness</td>
<td>429.360 (11.141)</td>
<td>423.438 (14.844)</td>
<td>6.768</td>
<td>0.029</td>
</tr>
<tr>
<td>Incidence of causal verbs</td>
<td>12.375 (6.964)</td>
<td>8.910 (8.857)</td>
<td>4.123</td>
<td>0.016</td>
</tr>
</tbody>
</table>

\(^a\) Welch’s \( F \) value.

Table 4
Means (standard deviations), \( F \) values, and effect sizes for English and Spanish essays.

Table 5
Accuracy of discriminant function analyses.

<table>
<thead>
<tr>
<th>Number of predictors</th>
<th>Accuracy of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>78.11</td>
</tr>
<tr>
<td>9</td>
<td>77.60</td>
</tr>
<tr>
<td>8</td>
<td>78.12</td>
</tr>
<tr>
<td>7</td>
<td>79.10</td>
</tr>
<tr>
<td>6</td>
<td>76.68</td>
</tr>
<tr>
<td>5</td>
<td>77.30</td>
</tr>
<tr>
<td>4</td>
<td>68.36</td>
</tr>
<tr>
<td>3</td>
<td>68.91</td>
</tr>
<tr>
<td>2</td>
<td>67.40</td>
</tr>
<tr>
<td>1</td>
<td>64.03</td>
</tr>
</tbody>
</table>
was most predictive of text authorship. As is typical of studies using discriminant function analyses, we describe the findings by reporting an estimation of the accuracy of the analysis (e.g., McCarthy et al., 2007). This estimation is made by plotting the correspondence between the actual texts (either L1 or L2) in the testing and training sets and the predictions made by the discriminant analysis (see Table 6).

The results show that the discriminant analysis, using the seven variables, correctly allocated 159 of the 201 essays in the training set (df = 1, n = 201) $\chi^2 = 67.99, p < .001$). For the test set, the discriminant analysis correctly allocated 159 of the 202 essays (df = 1, n = 202) $\chi^2 = 66.56, p < .001$). The model provides 79% accuracy.

Following standard procedure, this study reports its results in terms of recall and precision. Recall scores are computed by tallying the number of hits over the number of hits plus misses. Precision is the number of correct predictions divided by the number of incorrect predictions. This distinction is important because if an algorithm predicted everything to be a member of a single group it would score 100% in terms of recall but could only do so by claiming members of the other group. If this happened, then the algorithm would score low in terms of precision. By reporting both values, we can better understand the accuracy of the model. The accuracy of the model for predicting L1 and L2 texts written in English can be found in Table 7. The accuracy of the model for the training set was .79. The accuracy for the test set was .79. These results provide further evidence that there are lexical differences between L1 and L2 writers.

**Discussion**

This study has demonstrated that deeper-level lexical indices related to cohesion and network models taken from the Coh-Metrix tool can significantly distinguish between L1 and L2 texts. In this section, we will discuss the advantages and disadvantages of Coh-Metrix, elaborate on our findings and their relation to previous research in L2 writing, and provide implications for writing theory and research.

This study has provided an overview of a powerful computational tool for L2 textual analysis, Coh-Metrix (Graesser et al., 2004). The tool itself is freely available on-line (http://cohmetrix.memphis.edu). We then demonstrated how the data Coh-Metrix reports can be used to classify L1 and L2 texts using powerful statistical analyses such as discriminant functional analyses. The obvious advantage of the Coh-Metrix tool is its ability to report on linguistic indices of text in a manner that is fully automated. Many of these indices are deeper-level measures of cohesion and linguistic proficiency (e.g., LSA values and WordNet scores) that cannot be measured objectively by human raters. Thus the measures are necessarily computational in nature. Those indices that could be measured by

<table>
<thead>
<tr>
<th>Actual text type</th>
<th>Predicted text type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>83</td>
</tr>
<tr>
<td>Spanish</td>
<td>21</td>
</tr>
<tr>
<td>Test set</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>72</td>
</tr>
<tr>
<td>Spanish</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Actual text type</th>
<th>Predicted text type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>83</td>
</tr>
<tr>
<td>Spanish</td>
<td>76</td>
</tr>
<tr>
<td>Test set</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>72</td>
</tr>
<tr>
<td>Spanish</td>
<td>77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text set</th>
<th>Prediction</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.798</td>
<td>0.786</td>
<td>0.792</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.783</td>
<td>0.795</td>
<td>0.789</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text set</th>
<th>Prediction</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.788</td>
<td>0.786</td>
<td>0.787</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.786</td>
<td>0.788</td>
<td>0.787</td>
</tr>
</tbody>
</table>

Table 6
Predicted text type versus actual text type results from both training set and test set.

Table 7
Precision and recall finding (training and test set).
human raters are automated in Coh-Metrix to a degree that subjectivity and time limitations are not issues. These are not issues because Coh-Metrix reports standardized indices that do not vary within the same text (as human raters might) and reports these indices in rapid, automated manner. There are, however, disadvantages to the Coh-Metrix tool. Like all computational tools, it cannot consider context. Nor can Coh-Metrix measure rhetorical styles, truth values, and background knowledge.

However, the measures reported by Coh-Metrix can be used to elaborate on differences in the use of linguistic features between two populations. In the case of this study, our findings provide support for the idea that L1 writers provide a more cohesive writing product than L2 writers, and that the lexical choices employed by L1 writers demonstrate greater lexical associations. For instance, the hypernymy measure demonstrates that L1 writers create texts that are more abstract in their lexical choices, suggesting that L1 writers have access to higher conceptual hierarchies and stronger lexical networks. This provides support to notions that L2 learners have less developed lexical networks and overall proficiency, especially in reference to the hypernymic relationships contained within concepts (Crossley et al., 2009). The use of less abstract words is evident in some of the rhetorical strategies of L2 writers. Of particular interest is the listing of more specific hyponyms by L2 writers when providing examples. This trend was not noticed in L1 examples. L2 examples include:

[1] (L2): Are not all unbelievable advances (computers, chips, spaceships...) products of somebody’s imagination?
[2] (L2): Who is against belonging the EU—Sweden, Norway and Finland.
[3] (L2): It gives us a false image of the world full of marvelous and exciting cars, bodies, clothes.
[4] (L2): Money is behind the majority of murders, assassinations, robberies, assaults.

In consideration of polysemy values, the polysemy measure demonstrates that L1 writers use words that have more sense relations. This allows for the contention that L1 texts depend on words that have more meanings and provides evidence for wider lexical networks. Such notions have been discussed in previous research (Schmitt, 1998) following the hypothesis that L2 learners have less access to the same sense relations as L1 speakers. Given that the multiple senses found in a highly polysemous word are connected through a core sense (Murphy, 2004), we would argue that L1 learners have more developed lexical networks that allow for multiple sense retrieval. This can be seen in the following examples, where L1 writers use polysemous words (backbone, position, give, and produce) to support their arguments. Note that many of the examples are idiomatic.

[5] (L1): Religion is still considered the backbone of the world.
[6] (L1): Women are usually the least likely candidates for certain positions.
[7] (L1): Giving birth and producing a new life should not be a punishment in the working world.

In addition, L2 writers seem to create texts that depend less on “meaningful” words than L1 writers. While a conceptual measure in itself, word meaningfulness relates to how words have a priming effect that leads to the recall and activation of other words through lexical associations (Ellis & Beaton, 1993). This measure relates to the notion that L2 writers are less likely to use words that have stronger associations with other words and concepts. Words with higher meaningfulness scores are important for facilitating the comprehension of new vocabulary words and developing ideas that are not context dependent. Thus, this study provides evidence that L2 writers are less skilled at producing comprehensible texts that are context independent. Examples from the L1 and L2 essays are given below. In the first excerpt (L1), the words music, artist, and movies have high meaningfulness scores and are strongly associated with other lexical items. In the second excerpt (L2), the words fill and amount score low for meaningfulness and thus are not strongly associated with other words.

[8] (L1): Teenagers see music artists and movie stars doing drugs and assume that they too should do drugs.
[9] (L2): Since to all that you want to do you must fill in a great amount of papers.

L2 writers also appear less able to provide meaningful co-referentiality by using less given information in their writing as compared to L1 writers. Cohesive links such as overlap and givenness have been shown to construct a more cohesive writing product and aid in text comprehension (Hempelmann et al., 2005; Kintsch & Van Dijk, 1978). A lack of such links supports the notion that L2 writers are less likely to produce cohesive, readable texts than L1 writers. In
reference to givenness scores, L2 writers likely introduce too many new words and ideas. This is seen in the following L2 examples about the effects of television, where the first proposition and the second proposition of each demonstrate few given information links.

[10] (L2): The difference between the man and the animal is that the man can engage in a dialogue with other person, I mean, they can listen, analyze, argue and express an opinion. The television does these things instead of yourself, it gives you the reason without other possibility; but those reasons are not always true, even to the person who says them.

[11] (L2): The dialogue, characteristic of the person, requires an effort and it’s plain that, nowadays, people avoid doing something that requires effort; they look for comfort, not complications, they prefer to allow themselves to be carried. That is normal because he has a body that complains of the effort, but it is a pity because we are not only a body, we have intelligence and volition too, so we can see if something suits us and do it although our body asks us the opposite.

In reference to argument overlap, L2 writers appear to be less able to provide lexical overlap between propositions. Lexical overlap between propositions are important because overlapping vocabulary has been shown to aid in text comprehension and readability (Crossley et al., 2008; Kintsch & Van Dijk, 1978). The results of this study indicate that L2 writers do not provide readers with lexical overlap to the same degree as L1 writers and thus likely produce text that is less readable. This can be seen in the following L1 and L2 examples. The first example, from an L2 writer, appears to show little lexical overlap between the first and second sentence. The second example, from an L1 writer, seems to demonstrate greater lexical overlap between sentences.

[12] (L2): Dreams are the instruments of imagination, through them we begin to give shape to all the formless ideas that wander in our brain. These are such poetic notions that everybody will laugh at them, no doubt of it.

[13] (L1): In the past century, the world has made great strides to advance science, technology, and industry. For example technology has gone from Edison’s light bulb to home computers that calculate thousands of variables in a matter of minutes. With these great advancements in science, technology and industry, people must now compete with everyone else around the world, not just with everyone in their own countries.

In addition, L2 writers appear to create a less coherent situation model for the reader by providing less spatial and causal cohesion. Both causality and spatiality help ensure that the situation model of the text is well-constructed (Zwaan, Langston, & Graesser, 1995), which allows for better transfer of text meaning. Causality by itself is also important for constructing relations between events and actions and to show causal relations between clauses. Texts with less spatial and causal cohesion likely provide the reader with fewer linguistic features with which to build coherent textual representations. This could influence text processing and comprehension. Examples of L1 use of causal relations can be seen in excerpts [14] and [15], which include the use of causal particles (thus and therefore) and causal verbs (aid and make). Examples of L1 use of motion verbs related to spatiality can be found in excerpt [16] (advancing and booming).

[14] (L1): Therefore, feminists aid in the cause.
[15] (L1): Thus, it is impossible to make a truly informed decision.
[16] (L1): With technology advancing more rapidly than ever before, industrialization booming, and science taking leaps and bounds beyond what we know already.

L2 writers did, however, use more locational nouns than L1 writers. This might be a result of locational nouns having a more concrete reference as compared to motion verbs or the result of L2 writers of English depending more on nouns than verbs as means of conveying intention (McClure, 1991). Many of the locational nouns found in the data were geographical (e.g., Brussels, Spain, Germany), although many were not (e.g., home, outside, office).

L2 writers also used more frequent words than L1 writers. Frequent words, while allowing for easier comprehension, are likely not as well suited for creating the semantically complex and sophisticated arguments that mature readers expect. Thus, L2 writing which features more frequent words might actually be thought of as less proficient because the reader expects more lexical sophistication (Grant & Ginther, 2000; Reppen, 1994).
Of interest is the contradictory finding that L2 writers use more frequent words, but also use words that have a higher age of acquisition score. This could be explained by noting that the CELEX frequency measure chosen in this study relates to written words, while the age of acquisition score relates to spoken lexical production. To test this hypothesis, a simple correlation was conducted to assess whether the two measures were positively related. A Pearson’s correlation found that the two measures were negatively correlated \( (N = 201, r = -0.193, p < 0.01) \), lending credence to the idea that these two indices measure different phenomena. Thus, it is possible that L2 writers depend on more writer-based prose than reader-based prose as they depend on frequent written words, but more advanced spoken words when writing. This line of reasoning is in line with Reynolds’ (2002) contention that no transition from oral to written usage of causal connectors exists for L2 writers. This also implies that the L1 writers in this study likely use less frequent written words and do not depend on common spoken words as often as L2 writers.

Taken together, these findings present a portrait of L2 writers as lexically less proficient than L1 writers and provide additional support to the notion that L2 writers provide less lexical variation and sophistication in their writing (Linnarud, 1986; Silva, 1993). While similar findings have been documented in the past, our findings highlight the importance of lexical acquisition and its relation to writing proficiency. They also highlight differences between L1 and L2 writers in relation to deeper-level lexical indices that have correlates in cohesion and network models.

There are many implications that can be drawn from this study for L2 writing and L2 lexical proficiency. Of primary importance is the notion that L2 writers differ in their use of cohesive devices as compared to L1 writers. This implies that L2 writers are less likely to create a coherent text that is as readable and thus as comprehensible as the text of an L1 writer. L2 texts are also less abstract and less ambiguous. As a result, L2 texts are more likely to be context dependent. Additionally, L2 writers appear to have less-connected lexical networks than L1 writers. Consequently, the word choices employed by L2 writers have fewer associations and relationships. As a result, L2 written texts are likely to be more lexically and semantically disengaged than L1 texts and provide readers with fewer prospects for making links between words.

All of these findings demonstrate important differences between L1 and L2 writing. These differences are critical because they involve the lexicon, which is strongly related to the production of global errors (de la Fuente, 2002; Ellis et al., 1994; Ellis, 1995) and academic achievement (Daller et al., 2003). Additionally, the study examines lexical features that are not found solely at the surface level and instead examines coherence relations and lexical knowledge based on deeper-level, linguistic features.

Conclusion

From a purely practical approach, this study is of interest as it allows us to better understand the linguistic differences between L1 and L2 writers of English. Additionally, unlike past studies, this study was able to successfully distinguish to a significant degree L1 and L2 texts based solely on lexical features. Methodologically, the study presents a powerful text analysis tool, Coh-Metrix to the L2 writing community. The study also provides a thorough overview of a robust statistical algorithm, discriminant analysis. Both should prove useful in analyzing second language writing samples in the future.

While this study has important implications for differences between L1 and L2 lexical proficiency, the use of computational tools, and statistical analyses, it does have limitations. First, the study only considers lexical features and thus it ignores other variables such as syntactic complexity and part of speech (POS) tag density, which are also important to cohesion. Future studies using the Coh-Metrix tool would benefit from using many of the other 660 indices available in the tool. These include indices related to syntactic complexity, POS density, connectives, temporality, and anaphoric resolution, to name a few. Second, this study only looks at two language populations (English and Spanish). Future studies should examine whether the lexical differences noted here are generalizable and based on proficiency or if language-specific differences exist. Third, a few of the indices selected from Coh-Metrix are not representative of the entire lexical knowledge contained within a language. WordNet, for instance, contains over 170,000 lexical items, but it is still incomplete. The MRC psycholinguistic database contains 98,538 entries, but ratings vary depending on the measure. This could also help explain the contradictory findings with the age-of-acquisition scores in that the database might not provide representative coverage of the English lexicon. This, however, warrants further research before conclusions can be supported. In addition, no independent measures of proficiency were employed for the L1 and L2 corpora. Future studies should attempt to evaluate texts using independent measures such as human ratings to examine variance between L1 and L2 writings. Lastly, while the findings presented in this
study were statistically significant, the reported effect sizes for some of the variables were relatively small. However, small effect sizes can still be theoretically important because they provide evidence about the operational processes of variables in the real world (Abelson, 1985). Small effect sizes have also proven to be important in cumulative studies and can have important implications in practical contexts (Prentice & Miller, 1992).

This study further advances our understanding of the true nature of L2 writing in that it provides empirical evidence as to the differences between L1 and L2 writing using sophisticated tools and statistical methods. Such an approach complements past research, but, importantly, provides a design for future studies. Specifically, we see the need for studies examining different language populations, studies that consider how well computational tools predict human judgments of text proficiency, and the possibility of using computational tools such as Coh-Metrix to evaluate L2 essays.

Acknowledgments

The research was supported in part by the Institute for Education Sciences (IES R3056020018-02 and R305A080589). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors are indebted to Dr. Philip McCarthy of the Institute for Intelligent Systems for his assistance with the statistical analyses reported in this paper.

References


