Shared features of L2 writing: Intergroup homogeneity and text classification

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

a Department of Applied Linguistics, Georgia State University, 34 Peachtree St. Suite 1200, One Park Tower Building, Atlanta, GA 30303, United States
b Psychology/Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, United States

Abstract

This study investigates intergroup homogeneity within high intermediate and advanced L2 writers of English from Czech, Finnish, German, and Spanish first language backgrounds. A variety of linguistic features related to lexical sophistication, syntactic complexity, and cohesion were used to compare texts written by L1 speakers of English to L2 writers of English in order to examine if L2 writing shares text similarities regardless of the L1 of the writer. The results of the study provide evidence for intergroup homogeneity in the linguistic patterns of L2 writers in that four word-based indices (hyponymy, polysemy, lexical diversity, and stem overlap) demonstrated similar patterns of occurrence in the L2 writer populations sampled. However, significant differences were reported for these indices between L1 and L2 writers. The results of this study provide evidence that some aspects of L2 writing may not be cultural or independent, but rather based on the amount and type of linguistic knowledge available to L2 learners as a result of language experience and learner proficiency level.

© 2011 Elsevier Inc. All rights reserved.

Keywords: Computational linguistics; Corpus linguistics; L2 writing; Working memory; Second language writing; Intergroup homogeneity

Introduction

A key research strand in studies of second language (L2) writing has been the influence of the writer’s first language (L1) (Kubota, 1998; Matsuda, 1997). Numerous studies have described the ways in which L2 writers’ L1 can influence their L2 written production (i.e., intergroup heterogeneity) (Connor, 1984; Jarvis, 2010; Johns, 1984; Reid, 1992; Scarcella, 1984; Ventola & Mauranen, 1991). These studies have generally supported the notion that transfer from the L1 to the L2 occurs in the recursive strategies the writers use (i.e., planning and brainstorming), the rhetorical structure of the text, and the linguistic features produced (i.e., the lexicon, syntactic constructions, and the use of cohesive devices) (Grabe & Kaplan, 1996). Fewer studies have investigated similarities among L2 writers regardless of language background (i.e., intergroup homogeneity) and the potential for such similarities to characterize L2 writing (e.g. Hinkel, 2002; Reid, 1992).

Like these previous studies (Hinkel, 2002; Reid, 1992), we are interested in examining the presence of linguistic features in essays produced by L2 writers in English that are shared among the writers regardless of their L1. However,
unlike Hinkel (2002), we directly investigate intergroup homogeneity within L2 writers of English and, unlike Reid (1992), we focus on a variety of linguistic factors, such as lexical sophistication, syntactic complexity, and cohesive devices, all of which are important indicators of writing knowledge and common areas of interest for studies concerning working memory processes (Kellogg, 1996; McCutchen, 1996, 2000; Schoonen, Snellings, Stevenson, & van Gelderen, 2009).

Our primary research goal is to investigate the notion that L2 writers produce texts that share similar linguistic features regardless of the L2 writer’s L1. In particular, we are interested in examining college level L2 writers’ use of linguistic factors as a shared intergroup construct. We use texts written by native speakers of English as baseline examples of English college level writing and compare the production of lexical, syntactic, and cohesive features in the native speaker samples to those of the L2 writers of English. Such an approach permits us to examine if L2 writers produce texts that differ predictably from a baseline population, as well as to analyze patterns in L2 writing as generalizable occurrences that are shared across a range of L2 writers. Such an approach is important because it affords an examination of linguistic features inclusive to L2 writers as a composite. If we can predict general linguistic features that characterize L2 writers, then we can better understand the unique nature of L2 writing (Cumming, 2001; Hedgcock, 2005). In addition, such an approach can provide evidence that some aspects of L2 writing may not be cultural or independent, but rather based on the amount and type of linguistic knowledge available to L2 learners as a result of language experience and learner level (i.e., related to work memory processing).

Second language writing processes

When exploring differences between L1 and L2 writing, researchers generally make a distinction between higher-order operations and lower-order operations. Higher order operations such as planning, brainstorming, and text evaluation play an important role in developing ideas and revising writing. Such factors link strongly to a writer’s L1 and thus contribute to explaining writing proficiency as a function of cross-linguistic influences (Cumming, 1990). Lower-order factors, on the other hand, are generally linguistic in nature and help L2 writers “transform the propositional content of the message into language” (Schoonen et al., 2009, p. 79) as well as structure their ideas into a pre-developed plan and later modify the ideas and the essay structure during the revision process. The process of selecting the appropriate words, structuring those words syntactically, and then ensuring that the words and the structure are cohesive is a major struggle for many L2 writers (Bell & Burnaby, 1984; Bialystok, 1978; Brown & Yule, 1983; Nunan, 1989; Schoonen et al., 2009; White, 1981) and one that is influenced by their first language (Jones & Tetroe, 1987).

Lower-order factors also affect the working memory resources of L2 writers (Kellogg, 1996; McCutchen, 1996, 2000; Schoonen et al., 2009) because the complexity of writing requires conscious attention to word choice, syntax, text connections, and text organization, all of which can overload the working memory of the writer, especially in reference to processing and storing linguistic items (McCutchen, 1996; Scardamalia, 1981). Many L1 writers have automatized the aspects of lower level writing skills such as the lexical and syntactic production needed for text generation. Thus, for these writers, writing can occur with little conscious attention to short-term working memory processes (McCutchen, 2000) and resources can instead be devoted to higher-level demands such as text organization (Ransdell & Levy, 1996). However, L2 writers are less likely to effectively manage the complexities of writing because they may not have automated lexical and syntactic resources and therefore must devote conscious attention to word and syntactic choices (Schoonen et al., 2009). This is especially true in reference to lexical retrieval, which demands much more attention in an L2 than in an L1 (Chenoweth & Hayes, 2001). Committing attention to lower level writing skills influences the ability of some L2 learners to attend to higher-level strategies such as metacognitive knowledge and tapping into past writing experiences (Schoonen et al., 2009; Weigle, 2005), which, in turn, can affect the quality of writing produced (Schoonen et al., 2009).

Thus, in order to be successful, L2 writers need to have quick access to a large number of L2 words, phrases, and syntactic structures, along with the knowledge of how to combine these linguistic elements into a coherent piece. When such access is unavailable because of differences in language proficiency between the L1 and the L2, writers are left with a limited number of choices. These include relying on their L1 to help fill in missing linguistic information or relying on their existing L2 linguistic knowledge. The former tactic seems common, with many studies reporting links between lower level factors and the writer’s L1 (Connor, 1984; McClure, 1991; Reid, 1992). However, some lower level factors also show strong tendencies to be generalizable across L2 writing in English regardless of the L1 of the writer. Thus, similarities in lower level factors may exist across L2 writers as a population (Hinkel, 2002; Reid, 1992).
Examining links between a writer’s L1 and L2 is likely the most common approach for analyzing L2 writing. Such an approach is deeply steeped in theories of cross-linguistic influence (CLI), which is defined as the influence of a person’s knowledge of one language on the use or knowledge of another language. Past research has demonstrated that CLI affects almost all areas of linguistic and communicative competence in L2 learners (Jarvis & Pavlenko, 2008) and that CLI is more prominent in novice writers as compared to advanced writers (Rinnert & Kobayashi, 2009). Common features transferred during the writing process include linguistic items (Matsuda, 1997) such as lexical, syntactic, and cohesive features. Lexically, L2 writers have been shown to produce more overgeneralizations, frequent words (Crossley & McNamara, 2009; McClure, 1991), and synonyms (Connor, 1984). Syntactically, studies have demonstrated that syntactic categories can be used to classify L1 and L2 speakers of English (Mayfield-Tomokiyu & Jones, 2001) and can help predict the L1 of L2 writers in English (Koppel, Schler, & Zigdon, 2005). Additional studies have noted syntactic transfer in adverbial placement, relative clauses, and cleft constructions (Jarvis & Pavlenko, 2008). From a cohesion perspective, evidence supports the notion that text cohesion in L2 writing is influenced by the linguistic and rhetorical patterns in a L2 writer’s L1 (Ferris, 1994; Reid, 1992).

While research in CLI has focused on differences between L1 writers and L2 writers from a specific language background, other research has examined L2 writers as a homogenous group (an equivalence approach). That is to say, CLI research seeks to demonstrate invariance within a group of writers that share a common L1 and variance between that group of writers and another group (usually native speakers). In contrast, equivalence research focuses on potential invariance among L2 writers as a whole (i.e., intergroup homogeneity).

Perhaps the best example of equivalence research is Reid’s (1992) study in which she examined differences in essays written in English by native speakers of Arabic, Chinese, Spanish, and English in order to determine if differences in the production of cohesive devices existed between and among the language backgrounds. Reid investigated the use of four features of cohesion: pronouns, conjunctions, subordinate conjunction openers, and prepositions. Overall, she found that L2 writers, regardless of their L1, produced a significantly greater number of pronouns and conjunctions, as well as fewer prepositions, when compared to L1 writers. However, there were no shared similarities among the L2 writers in the production of subordinate conjunction openers. Reid interpreted these results as providing evidence for general differences in L2 writing as compared to L1 writing and suggested such differences could influence writing quality and present opportunities for direct pedagogical intervention.

Hinkel (2002) carried out a similar investigation to Reid (1992), except Hinkel’s focus was not on invariance among L2 writers as a homogenous group, but variance between L2 writers from specific L1 backgrounds and L1 writers of English. However, the results of her study enabled generalizations about shared linguistic features that characterized L2 writing across a wide range of languages. In her study, Hinkel (2002) examined over 1400 academic essays written by native speakers of English and L2 learners of English whose L1s were Chinese, Japanese, Korean, Vietnamese, Indonesia, and Arabic. For each text, she computed incidence scores for linguistic features (semantic and lexical classes for nouns, verbs, adverbs, pronouns, nominalizations, gerunds, tense, and aspect), subordinate clause features (nouns, adjective, and adverb clauses), and rhetorical features (conjunctions, exemplification, hedges, and emphatics). She then compared the incidence of these linguistic features between language groupings of L2 essays and L1 essays. Hinkel’s (2002) analysis reported numerous features that distinguished L1 essays from specific groupings of L2 essays, such as vague nouns, private and public verbs, modal verbs, amplifiers, emphatics, and tense and aspect markings. When differences between specific language groups writing in English and L1 writers were collapsed, patterns emerged that allowed Hinkel (2002) to define L2 writing as generally being similar to personal narratives because both contain restricted syntactic variety and complexity and limited lexical sophistication.

Our goal in this study is to investigate the potential for linguistic features related to text cohesion, lexical sophistication, and syntactic complexity to discriminate between texts written by L1 and L2 writers. We are particularly interested in examining if general linguistic differences exist between L1 and L2 essays that are likely not specific to the first language of the L2 writer. Such an approach is different than previous studies (e.g., Koppel et al., 2005) because it focuses on intergroup homogeneity as compared to intergroup heterogeneity. In order to examine potential linguistic differences between L1 writers and L2 writers, we use the computational tool Coh-Metrix to analyze the linguistic features in a corpus of L1 and L2 essays. We use the results computed by Coh-Metrix as the foundation for a statistical analysis to discriminate between L1 and L2 essays. Because our interest is in general differences between L1 and L2 writers, our corpus consists of a sub-corpus of L1 essays and four L2 sub-corpora of essays written by English learners from four language backgrounds: Czech, Finnish, German, and Spanish.
Methods

Corpus selection

We selected L2 essays that were written by writers from a variety of L1 language backgrounds to ensure our findings were not the result of shared first language features and to test the generalizability of our findings. Our selected corpus consisted of four L1 backgrounds from four different language families: Czech (Slavic), Finnish (Finno-Ugric), German (Germanic), and Spanish (Italic). The essays were taken from the International Corpus of Learner English (ICLE). 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 to 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), level (high intermediate to advanced writers), and mother tongue. The ICLE 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 afford discourse-orientated as well as grammatical and lexical investigation.

A comparison corpus of L1 essays was also collected from undergraduate students at a large university in the United States. The L1 corpus was designed to closely follow the criteria used for the ICLE corpus, except for the second language criterion. The 211 essays that comprised the L1 corpus were collected from native English speaking college students in first-year persuasive writing classes. The students sampled were not from remedial or advanced classes. The essays collected were all argumentative essays based on 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. Like the majority of the ICLE essays, the essays were untimed essays written outside of the classroom. This means that referencing of outside sources was allowed, but was not necessary.

The linguistic features of the 211 L1 texts were used as a baseline with which to compare the L2 texts sampled from the ICLE corpus. Thus, the L1 texts provide us with internal validity in that they furnish a starting point from which to compare and contrast the L2 texts, allowing us to determine in what ways L2 writers of English differ from L1 writers in their production of linguistic features. Our methodology is similar to that used by Reid (1992) and Crossley and McNamara (2009). As was the case with these past studies, the L1 texts that form the basis for this analysis should be seen as a baseline for comparison, not an ideal.

Statistical analysis

To examine the hypothesis that linguistic features differentiate L1 texts from L2 texts, we conducted a Multiple Analysis of Variance (MANOVA) followed by a discriminant function analysis (DFA). The MANOVA was conducted to select variables that demonstrated significant differences between L1 and L2 texts. A DFA is a common approach used in many previous studies that distinguished text types (e.g., Biber, 1993; Crossley & McNamara, 2009) and was used in this study to analyze differences between the L1 essays and the L2 essays taken from the ICLE corpus in order to determine if linguistic features distinguished L2 texts from L1 texts. Follow-up analyses compared each individual L2 group to the L1 baseline and reported on the accuracy of the DFA to correctly classify the essays by language group.

For the statistical analysis, we selected Coh-Metrix indices that measured linguistic features related to areas of interest in L1 and L2 writing including cohesion (e.g., connectives, word overlap, semantic co-referentiality), lexical sophistication (e.g., word concreteness, word imagability, lexical diversity, and word frequency), and syntactic

Table 1
Most common essay topics in the ICLE.

<table>
<thead>
<tr>
<th>Prompt</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? 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. In his novel ‘Animal Farm’, George Orwell wrote “All men are equal: but some are more equal than others”. How true is this today? Feminists have done more harm to the cause of women than good.</td>
</tr>
</tbody>
</table>
complexity (e.g., syntactic similarity and number of words before the main verb). The indices selected for this study come from measures of lexical coreferentiality, semantic coreferentiality (Latent Semantic Analysis), word frequency, lexical diversity, word information from the MRC Psycholinguistic database, hypernymy and polysemy values from WordNet, spatial cohesion, causal cohesion, temporal cohesion, and syntactic complexity. These measures are briefly discussed below. More detailed information about these indices can be found in Graesser, McNamara, Louwerse, and Cai (2004) and Graesser and McNamara (in press). More detailed information about links between these indices and L1 and L2 writing can be found in Crossley and McNamara (2011, in press).

Lexical co-referentiality
Coh-Metrix considers four forms of lexical co-reference between sentences: noun overlap between sentences, argument overlap between sentences, stem overlap between sentences, and content word overlap between sentences.

Latent Semantic Analysis (LSA)
Coh-Metrix measures semantic coreferentiality using LSA, which is a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts (Landauer, McNamara, Dennis, & Kintsch, 2007). The LSA indices in Coh-Metrix also measure given/new information (Hemelmann et al., 2005), which relates to the amount of information available for recovery from the preceding discourse (Halliday, 1967).

Word frequency
Coh-Metrix measures how frequently words occur in the English language. The primary frequency count in Coh-Metrix comes from CELEX (Baayen, Piepenbrock, & Gulikers, 1995), 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.

Word information (MRC Psycholinguistic Database)
Coh-Metrix calculates word information using 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. MRC word familiarity, concreteness, imagability, meaningfulness, and age of acquisition scores measure lexical constructs such as spoken word exposure (familiarity), word abstractness (concreteness), the evocation of mental and sensory images (imagability), word associations (meaningfulness), and intuited order of lexical acquisition (age of acquisition).

Hypernymy and polysemy indices
Coh-Metrix measures the relative ambiguity of words by calculating their polysemy value, which refers to the number of meanings or senses a word contains. Coh-Metrix measures the relative specificity of a text by calculating word hypernymy values, which refer 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).

Spatial cohesion
Coh-Metrix represents motional spatiality through measurements of motion verbs and locational spatiality through measurements of location nouns (Dufty, Graesser, Lightman, Crossley, & McNamara, 2006). In Coh-Metrix, classifications for both motion verbs and location nouns are taken from WordNet (Fellbaum, 1998).

Temporal cohesion
Temporal cohesion is measured in Coh-Metrix in three ways: aspect repetition, tense repetition, and the combination of aspect and tense repetition.

---

2 A lower hypernymy score relates to less specific words, while a higher score relates to more specific words.
Table 2
Descriptive statistics for the L1 corpus, the combined L2 corpus, and the grouped language corpora.

<table>
<thead>
<tr>
<th>Language</th>
<th>Mean number of words</th>
<th>Standard deviation</th>
<th>Texts in training set</th>
<th>Texts in test set</th>
<th>Total corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>719.545</td>
<td>132.876</td>
<td>144</td>
<td>67</td>
<td>211</td>
</tr>
<tr>
<td>L2</td>
<td>668.333</td>
<td>274.621</td>
<td>599</td>
<td>304</td>
<td>904</td>
</tr>
<tr>
<td>Czech</td>
<td>865.021</td>
<td>249.975</td>
<td>123</td>
<td>60</td>
<td>183</td>
</tr>
<tr>
<td>Finnish</td>
<td>739.947</td>
<td>250.384</td>
<td>147</td>
<td>82</td>
<td>229</td>
</tr>
<tr>
<td>German</td>
<td>514.158</td>
<td>259.070</td>
<td>194</td>
<td>102</td>
<td>296</td>
</tr>
<tr>
<td>Spanish</td>
<td>640.528</td>
<td>189.211</td>
<td>134</td>
<td>61</td>
<td>195</td>
</tr>
</tbody>
</table>

Causal cohesion
Causal cohesion is measured in Coh-Metrix by calculating the ratio of causal verbs to causal particles (Dufty, Hempelmann, Graesser, Cai, & McNamara, 2005), which relates to the conveyance of causal content. The causal verb count is based on the number of main causal verbs identified through WordNet (Fellbaum, 1998; Miller et al., 1990).

Lexical diversity
Coh-Metrix measures lexical diversity using a variety of indices that demonstrate small text length effects (McCarthy & Jarvis, 2010). These indices include the Measure of Textual, Lexical Diversity (MTLD) (McCarthy & Jarvis, 2010), D (Malvern, Richards, Chipere, & Duran, 2004), and M (Maas, 1972).³

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. Second, there is an index that measures the mean number of high level constituents (sentences and embedded sentence constituents) per words in sentences. Lastly, there is an index that assesses syntactic similarity by measuring the uniformity and consistency of syntactic constructions in the text using phrasal and syntactic categories.

Variable selection
To select the variables from the chosen Coh-Metrix indices, we randomly divided the essays in the corpora into two groups based on a 67/33 split. The groups functioned as a training set (the 67% split) and a test set (the 33% split). The purpose of the training set was to identify which of the variables contained within the selected Coh-Metrix measures best distinguished the L1 and the L2 essays. We selected these variables through a MANOVA and used pairwise comparisons to ensure the reported effect sizes were the result of shared variance between the L2 language groups (i.e., the Czech, Finnish, German, and Spanish groups) and not the result of only one or two L2 language groups. The selected variables were later used to predict the L1 and the L2 essays in the training set using a DFA. The L1 and L2 essays in the test set data were later categorized using the reported co-efficients in the initial DFA model. Descriptive statistics for the L1 and L2 corpora as a group and by individual L2 languages are located in Table 2.

Results
MANOVA
A MANOVA was conducted using the selected Coh-Metrix measures as the dependent variables and the essays from the training set as the independent variables (L1 essays and L2 essays). We selected the variable with the largest effect size as the representative variable for that measure. We selected more than one variable from the MRC Database and WordNet because the indices measured different linguistic features.

³ Maas, as reported by Coh-Metrix, is reverse-scaled so that lower numbers indicated greater lexical diversity.
Before final selection of variables, we assessed collinearity between variables so as not to waste potential model power. In testing for collinearity, we ensured that no index pair correlated above \( r > .70 \) and that each variable passed tolerance tests (i.e., VIF and tolerance values). Pearson correlations revealed that the LSA Given/New index was highly correlated with the other LSA values. Because the LSA Given/New index reported higher effect values than other LSA indices, it was retained in the analysis. In addition, syntactic similarity indices were also highly correlated to indices of causality. Because causality reported higher effect values than syntactic similarity, it was retained in the analysis. Lastly, VIF and tolerance values were above acceptable criteria for word frequency indices. Thus, word frequency indices were not included in this analysis. Descriptive statistics for the final variables for this analysis are presented in Table 3. The variables are ordered based on effect size.

**Pairwise comparisons**

We conducted a series of pairwise \( t \)-test comparisons with LSD corrections to examine if significant differences existed between each of the language groups and the L1 essays for each of the selected linguistic indices. Such an analysis examines the generalizability of each linguistic feature to demonstrate differences between L1 and L2 essays for all the language groups analyzed in this study. The mean and standard deviation data from the pairwise analysis is located in Table 4.

**Lexical diversity**

Significant differences were noted between all the L2 groups and the L1 group in lexical diversity scores. The results demonstrate that L1 writers use significantly less lexical diversity than L2 writers regardless of first language.

<table>
<thead>
<tr>
<th>Variables</th>
<th>L1 essays</th>
<th>L2 essays</th>
<th>( F(1,472) )</th>
<th>( \eta^2_p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical diversity M</td>
<td>0.022 (.002)</td>
<td>0.019 (.002)</td>
<td>140.865</td>
<td>0.170</td>
</tr>
<tr>
<td>Stem overlap adjacent sentences</td>
<td>0.576 (.153)</td>
<td>0.373 (.184)</td>
<td>139.915</td>
<td>0.100</td>
</tr>
<tr>
<td>LSA givenness</td>
<td>0.337 (.033)</td>
<td>0.305 (.046)</td>
<td>76.289</td>
<td>0.082</td>
</tr>
<tr>
<td>Average word polysemy</td>
<td>4.188 (.357)</td>
<td>3.941 (.450)</td>
<td>60.845</td>
<td>0.081</td>
</tr>
<tr>
<td>Average word hypernymy</td>
<td>1.707 (.206)</td>
<td>1.569 (.192)</td>
<td>60.412</td>
<td>0.027</td>
</tr>
<tr>
<td>Word meaningfulness every word</td>
<td>355.094 (8.143)</td>
<td>350.948 (11.993)</td>
<td>18.626</td>
<td>0.024</td>
</tr>
<tr>
<td>Tense and aspect repetition</td>
<td>0.759 (.083)</td>
<td>0.794 (.084)</td>
<td>17.089</td>
<td>0.021</td>
</tr>
<tr>
<td>Causal verbs and particles</td>
<td>39.810 (9.843)</td>
<td>36.451 (16.662)</td>
<td>14.969</td>
<td>0.017</td>
</tr>
<tr>
<td>Number of locational prepositions and nouns</td>
<td>0.696 (.119)</td>
<td>0.656 (.120)</td>
<td>11.945</td>
<td>0.015</td>
</tr>
<tr>
<td>Incidence of negation connectives</td>
<td>13.361 (6.280)</td>
<td>15.777 (7.831)</td>
<td>10.278</td>
<td>0.012</td>
</tr>
<tr>
<td>Word familiarity content words</td>
<td>578.835 (5.304)</td>
<td>577.203 (6.549)</td>
<td>8.367</td>
<td>0.006</td>
</tr>
<tr>
<td>Number of words before main verb</td>
<td>4.663 (1.413)</td>
<td>4.296 (1.595)</td>
<td>4.184</td>
<td>0.006</td>
</tr>
<tr>
<td>Word imagability content words</td>
<td>392.342 (19.981)</td>
<td>396.483 (20.939)</td>
<td>3.898</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 3

Means (standard deviations), \( F \) values, and effect sizes (\( \eta^2_p \)) for L1 and L2 essays in training set.

<table>
<thead>
<tr>
<th>Variables</th>
<th>English essays</th>
<th>Finnish essays</th>
<th>German essays</th>
<th>Czech essays</th>
<th>Spanish essays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical diversity M</td>
<td>0.022 (.003)</td>
<td>0.019 (.002)</td>
<td>0.018 (.003)</td>
<td>0.020 (.003)</td>
<td>0.020 (.003)</td>
</tr>
<tr>
<td>Word meaningfulness every word</td>
<td>355.094 (8.143)</td>
<td>347.600 (10.488)</td>
<td>354.975 (12.721)</td>
<td>354.130 (10.826)</td>
<td>345.838 (10.516)</td>
</tr>
<tr>
<td>Average word hypernymy</td>
<td>1.707 (.206)</td>
<td>1.585 (.173)</td>
<td>1.541 (.217)</td>
<td>1.588 (.172)</td>
<td>1.576 (.188)</td>
</tr>
<tr>
<td>Average word polysemy</td>
<td>4.188 (.358)</td>
<td>4.028 (.306)</td>
<td>3.861 (.396)</td>
<td>4.011 (.705)</td>
<td>3.897 (.312)</td>
</tr>
<tr>
<td>Stem overlap adjacent sentences</td>
<td>0.576 (.153)</td>
<td>0.423 (.166)</td>
<td>0.299 (.195)</td>
<td>0.346 (.147)</td>
<td>0.450 (.174)</td>
</tr>
<tr>
<td>Number of words before main verb</td>
<td>4.663 (1.412)</td>
<td>4.255 (1.027)</td>
<td>4.613 (1.948)</td>
<td>3.444 (1.093)</td>
<td>4.662 (1.641)</td>
</tr>
<tr>
<td>Word familiarity content words</td>
<td>578.835 (5.304)</td>
<td>576.439 (5.794)</td>
<td>577.771 (7.240)</td>
<td>578.356 (5.626)</td>
<td>576.155 (6.860)</td>
</tr>
<tr>
<td>Tense and aspect repetition</td>
<td>0.759 (.083)</td>
<td>0.774 (.072)</td>
<td>0.813 (.091)</td>
<td>0.807 (.066)</td>
<td>0.777 (.092)</td>
</tr>
</tbody>
</table>

Table 4

Means (standard deviations) L1 essays and L2 essays grouped by language background in training set.
**Word meaningfulness**

Significant differences were reported between the Finnish and Spanish L2 groups and the L1 group in terms of word meaningfulness. No differences were reported for the Czech and German groups. This finding demonstrates that L1 writers generally use more meaningful words than L2 writers.

**Word hypernymy**

Significant differences were found between all the L2 groups and the L1 group for word hypernymy values. The results of the hypernymy analysis reveal that L2 writers use significantly less specific words than L1 writers.

**Word polysemy**

Significant differences were reported between all the L2 groups and the L1 group for word polysemy values. The results demonstrate that L1 writers use more words with multiple senses than L2 writers of English.

**Word imagability**

Significant differences were noted between the Finnish and German groups and the L1 group in reference to word imagability. Results for the Spanish group approached significance, but significant differences were not reported for the Czech group. The results show that German essays contain more imagable words than L1 essays while Finnish essays contain less imagable words.

**Incidence of negations**

Significant differences were found between the Finnish and Czech groups and the L1 group in terms of the number of negations the essays contained. No differences were noted for the German and Spanish groups. This finding shows that L1 essays contain significantly fewer negations than Finnish and Czech essays.

**Stem overlap**

Significant differences were noted for stem overlap between the L2 groups and the L1 group. The findings demonstrate that L1 essays contain more stem overlap than L2 essays.

**Number of words before main verb**

Significant differences were found between the Czech and Finnish groups and L1 group in the number of words before the main verb. No differences were reported for the German and Spanish groups. These findings demonstrate that L1 essays contain sentences that have significantly more words before the main verb than Czech and Finnish essays.

**Word familiarity**

Significant differences for word familiarity were found between the Finnish and Spanish groups and the L1 group. No differences were noted between the German and Czech groups. The results show that L1 essays contain more familiar words than do Finnish and Spanish essays.

**Tense and aspect repetition**

Significant findings were found between the German and Czech groups and the L1 group in reference to tense and aspect repetition. No differences were reported for the Spanish and Finnish groups. The findings suggest that German and Czech essays contain more tense and aspect repetition than L1 essays.

**Discriminant function analysis**

We conducted a discriminant function analysis that analyzed only the linguistic indices that demonstrated significant differences between all L2 language groupings and the L1 group. The purpose of this DFA was to examine the potential for only those indices that strongly characterize L2 writing samples from all the sampled L1 backgrounds to classify L1 and L2 essays. The indices included in this analysis were word hypernymy, word polysemy, lexical diversity, and stem overlap.

A discriminant analysis is a statistical procedure that is able to predict group membership (the L1 and L2 essays) using a series of independent variables (the selected Coh-Metrix indices). Unlike the MANOVA analysis, the DFA provides an
estimate of relative importance of each of the indices to separate the language backgrounds when examined simultaneously (Field, 2005; Meyers, Gamst, & Guarino, 2006). In this study, we are interested in three features of the discriminant analysis: The Wilks’s Lambda, the classification results, and the discriminant function co-efficients. The Wilks’s Lambda tests whether the function reported by the DFA is significant. The classification result assigns each essay into one of the two groups (L1 essays or L2 essays). The discriminant function coefficients demonstrate the contribution that each linguistic index makes in predicting the dependent variable (whether the essay was written by an L1 or L2 writer). We use the training set to generate a discriminant function. Later, we use the discriminant function analysis model from the training set to predict group membership of the essays in the test set. If the results of the discriminant analyses 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).

The stepwise discriminant function analysis retained all four indices. The Wilks’s Lambda for the function was significant, \( \Lambda = .705, X^2 (10) = 258.179, p < .001 \). The classification results for the DFA correctly classified 79.3\% of the essays as L1 or L2 essays, \( X^2 (1) = 182.267, p < .001 \). The reported Kappa = .469 indicated a medium agreement between the actual essay classification and the predicted essay classification for the training set. The classification results for the test set correctly classified 79.6\% of the essays as L1 or L2 essays, \( X^2 (1) = 91.547, p < .001 \). The reported Kappa = .464 indicated a medium agreement between the actual essay classification and the predicted essay classification for the test set. All results were well above chance (chance is 50\% for the two groups; see Table 5 for classification results). These results demonstrate that the four variables that were characteristic of the L2 writing samples, regardless of the L1 background, were able to successfully classify 80\% of the essays as being written by an L1 or an L2 writer.

The discriminant function coefficients (DFC) for the discriminant analysis correspond to the partial contributions of each variable in the discriminant function. The DFC from the discriminant analysis demonstrated that lexical diversity (DFC = .636) contributed the most to separating the groups. This index was followed by word hypernymy (DFC = .507), stem overlap (DFC = .330), and word polysemy (DFC = .323). Thus, the four selected indices all meaningfully contributed to discriminating between the L1 and L2 essays.

As is also common, 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 + misses (i.e., the number of essays classified correctly as L1 essays divided by the total number of essays classified as L1 essays). Precision is the number of correct predictions divided by the sum of the number of correct predictions and false positives (i.e., the number of essays classified correctly as L1 essays divided by the number of essay classified correctly as L1 essays plus the number of L1 essays incorrectly classified as L2 essays). 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 overall accuracy of the model. The overall accuracy of the model for the training set was .729. The overall accuracy for the test set was .726 (see Table 6 for accuracy scores). The results provide evidence that the four selected linguistic features can be used to classify L1 and L2 essays.

### Classification accuracy based on first language

We are also interested in the accuracy with which the model classifies essays based on the first language of the authors. To assess this accuracy, we calculated cross-tabulations using the DFA results for the training and test set, but

<table>
<thead>
<tr>
<th>Actual text type</th>
<th>Predicted text type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
</tr>
<tr>
<td><strong>Training set</strong></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>114</td>
</tr>
<tr>
<td>L2</td>
<td>124</td>
</tr>
<tr>
<td><strong>Test set</strong></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>54</td>
</tr>
<tr>
<td>L2</td>
<td>63</td>
</tr>
</tbody>
</table>

Table 5
Predicted text type versus actual text type results from both training set and test set (L1 and L2 corpus with four indices).
based on the first language of the essay writer and not the generic category of L2 essays. For the test set, the model reported the best accuracy for the German essays (96% accuracy). The lowest results were for the Czech essays (65%). Accuracy results for all L2 essays based on language grouping (along with the L1 essays) are located in Table 7. All results were well above chance (chance is 20% for the five groups).

**Discussion**

This study has provided evidence that intergroup homogeneity exists in the linguistic patterns of L2 writers. For the L2 essays sampled in this study, four linguistic features (hyponymy, polysemy, stem overlap, and lexical diversity) reported similar values across the L2 writers regardless of the writers’ L1. Importantly, these values were significantly different from the values reported for the L1 essays. In summary, the results demonstrate that L2 writers use less sophisticated lexical features (i.e., more generalizable words that are less ambiguous) and less sophisticated morphological features (i.e., less stem overlap) than L1 writers. In contrast, the trend reported for our lexical diversity index patterns counter to expectations and may indicate differences in general rhetorical strategies (i.e., stylistic and structural choices) between L1 and L2 writers. The four indices that separate L1 and L2 essays are telling because they are lexical and morphological in nature and thus likely reveal differences between L1 and L2 writers in the amount and type of linguistic knowledge readily accessible through automatic processes, especially in terms of lexical retrieval (Chenoweth & Hayes, 2001). Together, these four linguistic features are robust enough to detect L2 writers from L1 writers with an accuracy of over 70%. While such accuracy is well above chance and statistically significant, it should not be interpreted as demonstrating universal characteristics of L2 writers, but rather characteristics that are predictive of L2 writers and generalizable to the L2 writers in this context.

From a lexical perspective, L2 writers produce words that are less sophisticated in that the words are more generalizable (lower hyponymy scores) and less ambiguous (lower polysemy scores). Such a finding supports the notion that the lexicon is a primary concern of L2 writers in reference to text generation (Porte, 1996, 1997). Lexical generation is difficult for L2 learners because their lexical access and retrieval processes are less automatic than those for L1 writers (Schoonen et al., 2009). In the absence of automatic lexical processing, L2 writers appear more likely to use less-specific words (hyponyms) that entail a variety of general concepts. Similar lexical features are common in child language acquisition in that the words children first produce are semantically general (i.e., light verbs such as go, do, make, come) (Clark, 1978). Such words are first to appear because they are salient, frequent, and semantically appropriate (Ninio, 1999). However, with time, L1 learners replace these general words with more specific words causing the general words, while still productive, to become less frequent (Clark, 1978). Similar movements toward more specific words as a function of time learning a second language also occur for L2 learners (Ellis & Ferreira-Junior, 2009). Driving this word production is likely the notion that L2 learners have less developed hierarchical connections between related words (Crossley, Salsbury, & McNamara, 2009). As a result, L2 writers may produce

---

**Table 6**

Precision and recall finding (training and test set): L1 and L2 corpus with four indices.

<table>
<thead>
<tr>
<th>Text set</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training set</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>0.479</td>
<td>0.792</td>
<td>0.597</td>
</tr>
<tr>
<td>L2</td>
<td>0.941</td>
<td>0.793</td>
<td>0.861</td>
</tr>
<tr>
<td><strong>Test set</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>0.462</td>
<td>0.806</td>
<td>0.587</td>
</tr>
<tr>
<td>L2</td>
<td>0.949</td>
<td>0.793</td>
<td>0.864</td>
</tr>
</tbody>
</table>

**Table 7**

Accuracy of classifying text based on first language (four indices).

<table>
<thead>
<tr>
<th>First language</th>
<th>English</th>
<th>Czech</th>
<th>Finnish</th>
<th>German</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training set</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>0.799</td>
<td>0.756</td>
<td>0.735</td>
<td>0.913</td>
<td>0.724</td>
</tr>
<tr>
<td>L2</td>
<td>0.821</td>
<td>0.650</td>
<td>0.780</td>
<td>0.961</td>
<td>0.672</td>
</tr>
<tr>
<td><strong>Test set</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
less-specific words (hyponyms) when a more specific word (hyponym) is warranted. Thus, when the L2 writer cannot access a hyponym (e.g., sedan), they may instead produce a related hyponym (e.g., car). In addition, L2 writers likely do not have access to the same conceptual organization as L1 writers. Thus, while L1 writers’ conceptual organization will contain a lexical entry with most to all senses (frequent and infrequent) attributed to a word, L2 writers will likely have fewer senses per lexical entry and the strength of connections between the senses will be weaker (Crossley, Salsbury, & McNamara, 2010). Because of differences in conceptual organization, L2 learners are more likely to produce words with fewer senses because the meanings of these words and the strength of the connections between the words are more easily accessible. Overall, the amount or type of lexical knowledge available to L2 writers appears to press them into producing texts that are less specific and less ambiguous when compared to L1 texts.

In terms of overlap of ideas, L2 learners were less prone to produce words with shared stems between sentences (with stems defined as free root morphemes and bound root morphemes, but not pronouns). As a result, L2 essays have less referential cohesion between ideas in the essays producing texts that are likely less coherent for the reader. The reason L2 essays contained less stem overlap between sentences, as compared to L1 writers, is likely partially attributable to both linguistic and rhetorical differences. L1 writers, with a fluent morphology system, are able to maintain cohesion across sentences without depending solely on lexical repetition. As a result, L1 writers can rely on words that are related at the level of the stem (i.e., creative and creatively) to maintain connections across text segments. Linguistic strategies such as these may be less available to L2 writers who may not have a fluent morphology system in English that allows for the automatic production of morphologically related words. Rhetorically, L1 writers are also more likely to produce texts with greater word repetition. Such an assertion is supported by our lexical diversity findings, which indicate that L1 writers produce texts that have a lower lexical diversity than L2 writers. Traditionally, lexical diversity has been associated with lexical knowledge, with past studies demonstrating that more proficient L2 writers produce texts with greater lexical diversity (Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002). However, it is unlikely that the L2 writers sampled in this study were of higher lexical proficiency than the L1 writers. What then explains the differences in the lexical diversity values? As in the stem overlap analysis, the answer likely relates to rhetorical strategies associated with text cohesion. L1 writers of English in first-year composition classes seem to prefer a style of writing which depends on a greater repetition of key words, as compared to L2 writers, who focus less on word repetition and more on the variety of words produced. We hypothesize that first-year composition students focus more on essay cohesion, while L2 writers focus more on linguistic processes such as word choice. This finding may also relate to working memory processes because it demonstrates that L2 writers focus on local and not global features of text as a result of dedicating working memory to word searches and morphosyntactic considerations (Schoonen et al., 2009).

To help demonstrate the differences discussed above, we present excerpts from two essays written by an L1 and an L2 writer (Spanish L1). To save space, we focus specifically on the introduction and conclusion segments of the essays.

**Introduction paragraphs**

**L1 writer.** Dreams are human beings natural way to escape reality and imagine what could be in their lives. With all of the high tech gadgets that are created to help us escape from reality it seems that dreams and imagination aren’t important in the world today. One man’s dream will be vastly different from another mans, but that’s what makes people who they are, a man’s dreams are what makes him the man he is. Dreams are important for people in today’s world.

**L2 writer.** From my point of view the prison and the Justice System are outdated. But we should not rehabilitate criminals at all, because they are the scum of the humanity. They must be punished in a tremendous harder way.

**Conclusion paragraphs**

**L1 writer.** When society comes to a place where dreams aren’t important that will signal and end of the need of people. Dreaming will never cease to be important because it is such a personal thing to all people. Imagination is not something that can be taken away from people. Imagination keeps people occupied better than any product that will ever be created because it is an element of all humans’ nature dreams occupy the ideal moments in a person’s life. No person has ever gone threw life without and imagination. It just does not happen. Dreams are
what separate people and machines, dreams allow for growth, dreams allow for change. Not all dreams come full circle and are fulfilled but in the effort to fulfill dreams is when a persons life happens, dreams are the most important thing that people have because dreams change lives. Dreams will always be an important part of society.

*L2 writer:* To conclude: The best solution to eradicate problem is a military attack to the actual Government and Monarchy, because they are passive about the problem. They would be replaced by a tyranny presided by “Chiquito de la Calzada” and “El Fary”, who are strict Spanish rulers. They are the cruelest people I know, so, with they in the power no one would dare to commit a crime or a wicked act. Think about it. The world is in your hands. Fight of a perfect world!!

These paragraphs exemplify the overall differences reported in our statistical analysis. Specific differences in terms of hypernymy, polysemy, stem overlap, and lexical diversity values are presented in Table 8. While not easy to discern without the aid of computational algorithms, the L1 writer’s words are more specific than the L2 writers, especially in the production of words such as signal, end, occupy, circle, and dreams. The L2 writer, on the other hand, has a greater number of non-specific words such as humanity, way, problem, and act. From a polysemy perspective, the L2 writer’s words contain fewer senses. For instance, the words prison, humanity, and Spanish contain 2 senses each. The words outdated, monarchy, and punish contain just one sense each. On the other hand, the L1 essays contain numerous words with multiple senses. For instance, the words escape, create, world, important, personal, occupy, and growth have 5 or more senses. The words natural, life, end, away, and change have 10 or more senses.

Stem overlap and lexical diversity are easier to envision without the aid of computer indices. The stem overlap index that was most predictive was stem overlap between adjacent sentences. In the L1 essay samples, stem overlap between sentences is almost perfect with every single sentence except the fourth to last sentence in the conclusion paragraph (*It just does not happen*) sharing a word or multiple words with the previous sentence. The clearest stem that is shared throughout the sentences is dream. Almost every sentence in the two sampled paragraphs includes a shared stem such as dream, dreams, or dreaming. In contrast, the L2 essay contains only a single instance of stem overlap in the conclusion paragraph, where the final two sentences contain the stem world. The scarcity of overlapping stems is also a likely reason that lexical diversity differences exist between the samples. While lexical diversity indices will not capture stems that include derivational and inflection morphology changes, they will capture word overlap, including the use of pronouns. As with the stem overlap index, our lexical diversity index indicates that L1 writers have greater word repetition than do L2 writers. Again, this is exemplified in the sampled paragraphs by the repetition of words such as dream, man, people, escape, imagination, person, allow, and society. Such repetition of words is more infrequent in the L2 paragraph samples, with generally only closed class words such as articles and pronouns repeated.

**Conclusion**

Like Crossley and McNamara (2009), this study demonstrates the importance of lexical features in distinguishing L1 and L2 writing samples. Additional studies that attempt to replicate the findings of this study and assess the inclusion of additional linguistic and rhetorical structures that exhibit intergroup homogeneity are needed. Specifically, such analyses need to be extended to a greater number of language backgrounds, particularly to writers outside of the European continent, as well as to a greater variety of writing genres, to assess the generalizability of these findings outside of argumentative essays. Additionally, writing samples in follow-up studies should control for potential linguistic differences that result from topic (Hinkel, 2002) and writing proficiency.
Overall, this study has revealed that four linguistic features (hypernymy, polysemy, stem overlap, and lexical diversity) help to demonstrate intergroup homogeneity between four groups of L2 writers from different and disparate language backgrounds. Such a finding not only supports the notion that L2 writing, regardless of language background, shares similar features, but also that such features can be used to distinguish L2 writing samples from L1 writing samples. However, such a conclusion does not preclude the notion that L2 writing is also writer-independent as well as subject to cultural constraints (Grabe & Kaplan, 1996); it simply reflects that possibility that elements of L2 writing related to automatic processing and working memory are shared among L2 writers regardless of independent writing skills and cultural writing knowledge.

Acknowledgments

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

References


Scott Crossley is an Assistant Professor 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 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.