Predicting lexical proficiency in language learner texts using computational indices

Scott A. Crossley  
Georgia State University, USA

Tom Salsbury  
Washington State University, USA

Danielle S. McNamara  
University of Memphis, USA

Scott Jarvis  
Ohio University, USA

Abstract

The authors present a model of lexical proficiency based on lexical indices related to vocabulary size, depth of lexical knowledge, and accessibility to core lexical items. The lexical indices used in this study come from the computational tool Coh-Metrix and include word length scores, lexical diversity values, word frequency counts, hypernymy values, polysemy values, semantic coreferentiality, word meaningfulness, word concreteness, word imagability, and word familiarity. Human raters evaluated a corpus of 240 written texts using a standardized rubric of lexical proficiency. To ensure a variety of text levels, the corpus comprised 60 texts each from beginning, intermediate, and advanced second language (L2) adult English learners. The L2 texts were collected longitudinally from 10 English learners. In addition, 60 texts from native English speakers were collected. The holistic scores from the trained human raters were then correlated to a variety of lexical indices. The researchers found that lexical diversity, word hypernymy values and content word frequency explain 44% of the variance of the human evaluations of lexical proficiency in the examined writing samples. The findings represent an important step in the development of a model of lexical proficiency that incorporates both vocabulary size and depth of lexical knowledge features.

Keywords

computational linguistics, corpus linguistics, depth of lexical knowledge, hypernymy, lexical diversity, lexical frequency, lexical proficiency, vocabulary size

Corresponding author:
Scott A. Crossley, 34 Peachtree St. Suite 1200, One Park Tower Building, Atlanta, GA 30303
Email: sacrossley@gmail.com
Introduction

Although a clear definition of lexical proficiency is lacking, most research in second language (L2) studies indicates that lexical proficiency refers to vocabulary size, depth of vocabulary knowledge and the accessibility of core lexical items (Meara, 2005). The lexical proficiency of speakers, especially L2 learners, is of crucial interest to language acquisition and linguistic competence for three main reasons. First, misinterpretations of lexical items produced by learners are key elements in communication errors (Ellis, 1995); second, lexical proficiency strongly correlates with academic achievement (Daller et al., 2003); and third, understanding lexical acquisition in relation to its deeper, cognitive functions can lead to increased awareness of how learners process and produce language (Crossley et al., 2009; Crossley et al., 2010b; Salsbury et al., in press). Measuring lexical proficiency benefits from using assessments that evaluate more than just simple lexical knowledge, such as the number or variety of words used, the frequency of words used, or the ability to match dictionary definitions with strings of letters (Nation, 2005). Nevertheless, many studies of lexical proficiency have relied on these types of performance assessments (Polio, 2001), which provide only a partial and relatively surface level representation of lexical knowledge. A fuller understanding of lexical proficiency, along with a practical, computational model with which to assess lexical proficiency, would prove beneficial to language researchers, language learners, language teachers, and educational institutions.

The purpose of this study is to examine the potential for lexical indices to predict human evaluations of lexical proficiency based on lexical features related to vocabulary size, depth of knowledge, and access to core lexical items. Vocabulary size relates to how many words a learner knows (e.g. lexical features such as diversity or vocabulary size). Depth of knowledge features, in contrast, relate to how well a word is known (e.g. lexical features such as semantic relatedness, word sense relations, and word associations). Finally, access to core lexical items relates to how quickly words can be retrieved or processed (e.g. lexical features such as word concreteness and familiarity). Our study focuses on one simple question: what lexical indices are predictive of human ratings of lexical proficiency? To address this question, we test specific computerized indices against holistic human judgments of lexical proficiency in written texts. Our aim is to provide a formal model that helps to explain human judgments of lexical proficiency.

We are better able to address this question than our predecessors because developments in computational algorithms and models related to lexical processing now allow for the quick and accurate assessment of a variety of lexical features. These features can be measured using automated lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters, psycholinguistic word measurement indices, and other lexical components that have been developed in the fields of cognitive science, natural language processing, and computational linguistics (Jurafsky and Martin, 2008). These computational measures have the capability not only to measure vocabulary size features such as lexical diversity, but also depth of knowledge features such as semantic relatedness, word sense relations, and word associations, and access to core lexical items through psycholinguistic word properties. However, the examination of lexical knowledge, lexical output, and lexical development using these computational indices is still in
its infancy. Orchestrating these measures into a practical model from which to measure lexical proficiency has not occurred.

**L2 lexical assessment**

Most L2 assessment tests measure receptive skills (Pearson et al., 2007). Those tests that examine productive lexical knowledge have generally focused on breadth and depth of knowledge features. For instance, Lexical Frequency Profiles (LFP; Nation and Heatley, 1996) and P-Lex (Meara and Bell, 2001) both assess L2 lexical development using indices of lexical frequency (i.e. breadth of knowledge). Laufer and Nation (1995) argued that LFP could discriminate between L2 proficiency levels and that it correlated with independent measures of vocabulary knowledge. However, Meara (2005) criticized LFP for being less predictive, especially for shorter texts, than Laufer and Nation claimed. From a depth of knowledge perspective, Meara’s (2005) automated tool, V_Links, tests a learner’s vocabulary organization using a small number of words. Survey instruments are also common in assessing L2 learners’ depth of knowledge. Many of these instruments ask participants to generate word derivatives (Schmitt and Meara, 1997) or identify the word associates of a stimulus word (Read, 1993). One problem with survey instruments such as these is that testing depth of knowledge requires the extensive testing of individual words. Testing individual words is not only extremely time consuming, but also does not allow for the testing of a representative sample of words (Meara and Wolter, 2004). For instance, if a language learner was tested on only 50 words and information was extracted on 10 features of those words, a survey would need to be designed that tested 500 items.

**Measures of lexical proficiency**

Contemporary notions of lexical proficiency are premised not on the number of words a learner knows, but also on psycholinguistic models of word knowledge and lexical networks. Psycholinguistic word information relates to linguistic properties of words that affect word processing and learnability such as word concreteness, imagability, and familiarity. Lexical networks are the result of connections between conceptual levels, sense relations, semantic co-referentiality, and word associations (Crossley et al., 2009, 2010a, 2010b; Haastrup and Henriksen, 2000). In lexical networks, words interrelate with other words to form clusters of words that act categorically. These clusters connect to other clusters and other words, until entire lexicons are developed based on interconnections (Haastrup and Henriksen, 2000). Common lexical relationships that help form lexical networks include hierarchical relations such as hyponymy, sense relations such as those found in polysemy, semantic similarity connections, and word frequency effects. Measures related to lexical network models and psycholinguistic properties along with frequency and lexical diversity are discussed in detail below.

**Lexical diversity.** Historically, indices related to lexical diversity concentrated on type-token ratios (TTR), which are simple formulas that divide the number of different words (types) by the total number of words (tokens) in a given text. There are various transformations of
simple TTR measures such as Corrected TTR (Carrol, 1964), Log TTR (Herdan, 1960), D (Malvern et al., 2004), Advanced TTR (Daller et al., 2003) Guiraud Advanced (Daller et al., 2003), and the Measure of Textual Lexical Diversity (MTLD; McCarthy and Jarvis, in press). The premise behind lexical diversity indices is that more diverse vocabularies are indicative of more proficient and larger lexicons. Although lexical diversity indices are broad, versatile measures of lexical proficiency, they fail to account for important qualitative information about words, namely word difficulty (Vermeer, 2000; Daller et al., 2003).

**Lexical frequency.** Lexical frequency is often equated with lexical richness. Most frequency indices depend on frequency lists and are based on the hypothesis that a higher lexical proficiency results in the use of less frequent words (Meara and Bell, 2001). Word frequency has traditionally been assigned to the breadth of knowledge category, but this categorization is debatable. Ellis (2002), for instance, argues that the production and comprehension of words is a function of their frequency of occurrence in language. Under this approach, word frequency helps determine lexical acquisition because each repetition of a word strengthens the connections between the word and its meaning categorization. As learners are exposed to frequent words, there is a reduction in processing time because the practice time with the word increases. Such a model of lexical acquisition is supported by studies that demonstrate that high frequency words are named more quickly than low frequency words (Balota and Chumbly, 1984), are processed more quickly in reading tasks (Kirsner, 1994), are judged more quickly to be words in lexical decision tasks (Forster, 1976), and have faster response latencies (Glanzer and Ehrenreich, 1979).

Most studies concentrating on L2 reading and writing proficiency have found that beginning L2 learners are more likely to comprehend, process, and produce higher frequency words (Crossley and Salsbury, 2010; Ellis, 2002) and that beginning L2 learners use higher frequency words than advanced learners (Bell, 2003; Meara and Bell, 2001). However, recent studies into L2 lexical development that have focused on spoken data have demonstrated that beginning L2 learners might first produce words that are more concrete even if they are less frequent. Only later do learners appear to produce words that are more frequent. This is likely the result of frequent words having more senses and thus being more ambiguous in potential meaning (Crossley et al., 2010b).

**Word meaningfulness.** Word meaningfulness relates to how many associations a word has with other words (Toglia and Battig, 1978). Words with high meaningfulness include words like food, music, and people while words with low meaningfulness include acumen, cowl, and oblique. Words in the first list invoke multiple word associations, while those in the second list have fewer associations. It is argued that L2 learners have fewer and more varied word associations than native speakers (Schmitt and Meara, 1997; Zareva, 2007) because L2 learners have not developed close semantic links between words (Ellis and Beaton, 1993). Semantic links between words mediate the organization and storage of words. In addition, words with more associations are argued to be acquired first (Ellis and Beaton, 1993).
The link between word associations and lexical acquisition is supported in several recent studies. For instance, Zareva (2007) found that higher proficiency learners provide significantly more word associations than intermediate and beginning level learners. She argued that larger vocabularies allow for a greater number of word associations. A second, recent study (Salsbury et al., in press) explored word meaningfulness values provided by the MRC psycholinguistic database and compared the changes in the values as a function of time spent studying English. Salsbury et al. found that meaningfulness scores decreased as time spent studying English increased. The study suggests that higher proficiency learners use more difficult words that have fewer word associations. This finding was supported in Crossley and Salsbury (2010), who used computational indices to predict produced and not produced verbs and nouns in beginning L2 data. The results of this study demonstrated that word meaningfulness was a significant predictor of early verb and noun production in L2 learners. Overall, these studies support the notion that as a language learner’s network of word associations develops, a greater number of word associations become available to the learner. The strength of these associations allows learners to extend their lexical production to include peripheral words that have fewer associations.

**Word concreteness.** Word concreteness refers to here-and-now concepts, ideas, and things constituting core lexical knowledge (Toglia and Battig, 1978). On a continuum, very concrete words would be placed on one end and very abstract words on the opposite end. The concreteness of a word has implications for a word’s learnability because concrete words, as compared to abstract words, have advantages in tasks involving recall, word recognition, lexical decision, pronunciation, and comprehension (Gee et al., 1999). Regarding L2 lexical acquisition, studies have demonstrated that concrete words are learned earlier (Crossley et al., 2009; Salsbury et al., in press) and more easily than abstract words (Carter, 1987).

**Word familiarity.** Word familiarity is related to word frequency (Schmitt and Meara, 1997) in that words that are rated as more familiar tend to occur more frequently in text and discourse. However, familiar words are salient whereas frequent words may not be (e.g. determiners or prepositions). Few studies have investigated word familiarity directly. A recent study (Salsbury et al., in press) investigated the development of familiar words in L2 learners and found that L2 learners’ productive vocabulary did not significantly change in terms of word familiarity. Nonetheless, word familiarity remains an important feature of lexical proficiency to investigate because familiar words are likely core lexical items that act as anchors from which to extend word meanings and associations.

**Word imagability.** Imagability scores are important because a word or concept that triggers a mental image quickly and easily is more likely to be recalled. Thus, highly imagable words likely constitute core lexical items in a learner’s lexicon. Ellis and Beaton (1993) found that highly imagable words were good candidates for keyword techniques in second language vocabulary learning. Additionally, highly imagable words have more context availability (Schwanenflugel, 1991) because they are experienced and analyzed.
visually (Ellis and Beaton, 1993). Context availability and visual analysis facilitate the learning of words for L2 learners.

**Hypernymy.** Hypernymy is based on the connection between general and specific lexical items (Chaffin and Glass, 1990). Hypernymic relations are hierarchical associations between hypernyms (superordinate words) and hyponyms (subordinate words). A hypernym is a word that is more general than a related word (car as compared to convertible) and a hyponym is more specific than a related word (convertible as compared to car). Hypernymy is an important organizational system for lexical relations in network models (Haastrup and Henriksen, 2000; Crossley et al., 2009) because it allows for hierarchical categorizations that explain properties that hyponyms inherit from hypernyms. Thus, hypernymy affords the economical representation of lexical properties (Chaffin and Glass, 1990; Murphy, 2004) and lexical generalization (Murphy, 2004).

From a developmental perspective, hypernymic relations are more likely acquired as the learners advance cognitively (Anglin, 1993), as they increase their levels of education (Snow, 1990), and as they acquire more specific lexical knowledge (Wolter, 2001). L2 studies have demonstrated that learners use more words of general than of specific meanings (Levenston and Blum, 1977) and that L2 learners’ use of general terms often results in inappropriate overgeneralizations (Ijaz, 1986). More recent studies have demonstrated that hypernymic relations by L2 learners increase (become less specific) as time is spent studying English (Crossley et al., 2009). Hypernymy is also important in early L2 verb production (Crossley and Salsbury, 2010).

**Polysemy.** Polysemous words are words that have more than one related sense. For instance, the word *class* has at least six related senses, including socio-economic class, a body of students, a course of study, a collection of things sharing similar attributes, a league ranked by quality (usually sports related), and elegance in dress or behavior. Polysemous relations are connected to conceptual organization because words with multiple overlapping senses are thought to be part of the same conceptual structure (Murphy, 2004). Under a network approach, the multiple senses in a polysemous word are located in a single lexical entry. Such an approach suggests that separate entries for related word senses are uneconomical because they fail to capture the sense connections in the word’s uses. However, word senses located within a single lexical item allow learners to efficiently recognize meaning relationships between a word’s senses (Verspoor and Lowie, 2003).

Studies concerning the polysemy knowledge of L2 learners have found that word sense knowledge increases as L2 learners gain proficiency and that word senses are retained more than forgotten (Schmitt, 1998). Crossley et al. (2010b) also found that as learners’ language proficiency grew, their production of multiple word senses increased. Both studies provide evidence for the development of word senses in L2 learners.

In summary, past studies have demonstrated that lexical features related to vocabulary size, depth of knowledge, and access to core lexical items are important indicators of lexical growth and proficiency in L2 learners. Many of these features are important in supporting lexical network theories (e.g. hypernymy, word meaningfulness, and polysemy).
or measure psycholinguistic word properties of language (e.g. concreteness, imagability, and familiarity).

**Method**

Our goal is to determine the degree to which automated lexical indices related to vocabulary size, depth of knowledge, and access to core lexical items can predict human ratings of lexical proficiency. To accomplish this, we analyzed a corpus of scored writing samples using lexical indices taken from the computational tool Coh-Metrix (Graesser et al., 2004). To ensure a wide variety of lexical proficiency, we collected writing samples from ten L2 learners at the beginning, intermediate, and advanced levels, as classified through Institutional Test of English as a Foreign Language (TOEFL) scores. We also collected writing samples from a comparison corpus of native speakers. Trained raters scored the writing samples using a holistic lexical proficiency rubric. We first divided the scored writing samples into training and test sets (Whitten and Frank, 2005). To examine which lexical variables were most predictive of lexical proficiency, we conducted correlations and a linear regression model comparing the human lexical proficiency ratings and the Coh-Metrix lexical indices using the training set only. The results of this analysis were later extended using the regression model to the held back, independent test set data, and finally to the complete corpus.

**Corpus collection**

Writing samples were collected from participants involved in a longitudinal study in an intensive language program at a large university in the United States. The writing samples were unstructured and unprepared daily written journals (i.e., freewrites) that were completed as part of the learners’ regular intensive English coursework. The teachers did not edit or grade the journal entries, but they did comment on the content of the entries. These comments modeled for the learners correct word forms and specific vocabulary in the context of the students’ journal entry. The learners typically responded in writing to their teachers’ questions and comments in subsequent journal entries. All original texts were handwritten by the participants and later entered electronically by the researchers. The participants ranged in age from 18 to 27 years old and came from a variety of L1 backgrounds (Korean, Japanese, Arabic, French, Bambara, Portuguese, Spanish, and Turkish). Most of the participants began their intensive English studies in the first level of the 6-level program as determined by an in-house placement exam. The students also completed an institutional TOEFL exam every two months. One hundred and eighty L2 writing samples were selected from a corpus of written data collected from 10 L2 learners over a one-year period. The number of texts collected from each learner was not equal because all learners did not submit the same number of writing samples. All texts were corrected for spelling in order for the lexical items to be analyzed by Coh-Metrix. We did not correct lexical mistakes, though, because we are interested in lexical production, not accuracy.

To ensure appropriate lexical proficiency variability across the writing samples, Institutional TOEFL scores were used to classify the L2 writers’ samples into beginning,
intermediate, and advanced categories. Although classifications for level are not available for Institutional TOEFL scores, these scores have been compared to other proficiency measures that do include level classification. These include American Council on Teaching Foreign Languages (ACTFL) ratings (Boldt et al., 1992) and the TOEIC (Educational Testing Service). The classification criteria used in these proficiency measures were linked to the participants’ Institutional TOEFL scores to categorize the learners. Institutional TOEFL scores below 400 were categorized as beginning level; scores between 400 and 499 were categorized as intermediate level, and scores at 500 or above were categorized as advanced level. In total, 60 writing samples were selected for the three proficiency levels (beginning, intermediate, and advanced). However, not all students were represented at each level because not all learners entered the program at the beginning level and not all students progressed to the advanced level.

A comparison corpus of 60 native speaker free-writings was selected from the Stream of Consciousness Data Set from the Pennebaker Archive Project (Newman et al., 2008). The stream of consciousness corpus includes over 2300 free writing samples collected from freshman psychology students over a three-year period. Like the L2 writing samples, the free-writing samples were unstructured and unprepared. All writing samples collected from the native speaker corpus were classified under the category native speaker.

Thus, our writing sample corpus contained 60 writing samples from each proficiency level as well as 60 writing samples from the native speakers for a total of 240 sample texts. Writing samples were separated at the paragraph level. The samples were controlled for text length by randomly selecting a text segment from each sample that was about 140 words (depending on paragraph constraints; see Table 1 for more details). To ensure that text length differences did not exist across the levels, an Analysis of Variance (ANOVA) test was conducted. The ANOVA demonstrated no significant differences in text length between levels.

**Survey instrument**

The survey instrument used in this study was a holistic grading rubric that we adapted from the American Council on the Teaching of Foreign Languages’ (ACTFL) proficiency guidelines for speaking and writing (ACTFL Inc., 1999) and holistic writing proficiency rubrics produced by American College Testing (ACT) and the College Board (for use in SAT writing evaluation). The proficiency levels in these rubrics along with the evaluative sections related to lexical proficiency were combined to form an

<table>
<thead>
<tr>
<th>Learner type</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginner</td>
<td>135.583</td>
<td>32.300</td>
</tr>
<tr>
<td>Intermediate</td>
<td>145.417</td>
<td>26.029</td>
</tr>
<tr>
<td>Advanced</td>
<td>141.450</td>
<td>26.109</td>
</tr>
<tr>
<td>Native</td>
<td>140.850</td>
<td>14.769</td>
</tr>
</tbody>
</table>
outline from which to develop a holistic scoring metric to evaluate overall lexical proficiency. Five experts in language processing with Ph.D.s in either linguistics or cognitive psychology reviewed the rubric outline and made suggestions for the outline’s revision. Once these suggestions were implemented, the revised outline was then subjected to two usability tests: first by experts in the field of language processing and second by human raters (native speakers of English). The usability tests led to additional revisions and the final holistic rubric that we use in this study. The final rubric defined lexical proficiency as skillful language use that is accurate and fluent and is characterized by the appropriate use of conceptual categories, coherence, and lexical-semantic connections (see Appendix 1 for rubric).

**Human ratings**

To assess the 240 writing samples that comprise our written corpus, three native speakers of English were trained as expert raters. The three raters were all graduate students in an English Department at a large university in the United States. The raters were trained on an initial selection of 20 writing samples taken from a training corpus not included in the written corpus used in the study. The raters assigned each writing sample a score between 1 (minimum) and 5 (maximum). To assess inter-rater reliability, Pearson correlations were conducted between all possible pairs of rater responses. The resulting three correlations were averaged to provide a mean correlation between the raters. This correlation was then weighted based on the number of raters (Hatch and Lazaraton, 1991). If the correlations within the three raters did not exceed .70 \((p < .001)\), the raters were retrained on a new set of 20 writing samples. Using this approach, the raters needed to be retrained once. After the second training, the average correlation between the three raters was \(r = .796 \,(p < .001)\) with a weighted correlation of \(r = .921\).

**Variable selection**

All the lexical indices used in this study were provided by Coh-Metrix (Graesser et al., 2004). Coh-Metrix is a computational tool that reports on over 600 linguistic indices related to conceptual knowledge, cohesion, lexical difficulty, syntactic complexity, and simple incidence scores (see Appendix 2 for a screenshot of Coh-Metrix output). Many of the indices reported by Coh-Metrix are related to lexical proficiency. These indices can be separated into vocabulary size, depth of knowledge, and access to lexical core measures. The measures related to vocabulary size include indices of lexical diversity as reported by \(MTLD\) (McCarthy and Jarvis, in press) and \(D\) (Malvern et al., 2004). The measures related to depth of knowledge include indices for hyponymy, polysemy (as reported by WordNet; Fellbaum, 1998), semantic co-referentiality (as reported by Latent Semantic Analysis; Landauer et al., 2007), and word frequency (as reported by the CELEX database; Baayen et al., 1995). Those measures related to accessing core lexical units include word concreteness, word familiarity, word imagability, and word meaningfulness (as reported by the MRC Psycholinguistic Database; Wilson, 1988). We also include indices related to word length for comparison purposes. This provided us with a total of 10 measures. For information about the computation of these measures in
Coh-Metrix and their relation to theories of lexical proficiency, we refer the reader to Crossley and McNamara (2009), Crossley and Salsbury (2010), Crossley et al. (2009), Crossley et al. (2010a), Crossley et al. (2010b), and Salsbury et al. (in press).

To test our results on an independent corpus that was not used in the initial analysis, we used training and test sets. We divided the corpus into two sets: a training set \( (n = 180) \) and a testing set \( (n = 60) \) based on a 67/33 split. The purpose of the training set was to identify which of the Coh-Metrix variables best correlated with the human scores assigned to each freewrite. These variables were later used to predict the human scores in the training set using a linear regression model. Later, the freewrites in the test set were analyzed using the regression model from the training set to calculate the predictability of the variables in an independent corpus (Whitten and Frank, 2005).

In order to allow for a more reliable interpretation of the multiple regression, we ensured that there were at least 20 times more cases (texts) than variables (the lexical indices). A 20 to 1 ratio allows for the interpretation of each variable’s individual contribution to the regression model (Barcikowski and Stevens, 1975; Field, 2005). We used Pearson correlations to select a variable from each measure to be used in a multiple regression. Only those variables that demonstrated significant correlations with the human ratings were selected. With 160 ratings in the training set, this allowed us to choose 8 lexical variables out of the 10 selected. To check for multicollinearity, we conducted additional, correlation tests on the selected variables. If the variables did not exhibit collinearity \( (r < .70) \), they were then used in the multiple regression analysis. The selected measures and their respective indices are discussed below in reference to their importance in lexical proficiency.

**Results**

**Pearson correlations training set**

We selected the Coh-Metrix measures that demonstrated the highest Pearson correlation when compared to the human ratings of the written samples. The 10 selected variables and their measures along with their \( r \) values and \( p \) values are presented in Table 2, sorted by the strength of the correlation. No indices from either the polysemy measure or the word length measure demonstrated significant correlations. Thus only eight variables demonstrated significant correlations with the human ratings.

**Collinearity**

Pearson correlations demonstrated that the word concreteness score was highly correlated \( (> .70) \) with the word imagability score \( (N = 160, r = .935, p < .001) \). Because the word concreteness value had a lower correlation with the human ratings as compared to the word imagability score, the word concreteness value was dropped from the multiple regression analysis. Thus, there were only seven variables included in the final analysis.
A linear regression analysis was conducted for the seven remaining variables. These seven variables were regressed onto the raters' evaluations for the 160 writing samples in the training set. The variables were checked for outliers and multicollinearity. Coefficients were checked for both variance inflation factors (VIF) values and tolerance. All VIF values and tolerance levels were at about 1, indicating that the model data did not suffer from multicollinearity (Field, 2005).

The linear regression using the seven variables yielded a significant model, $F(3, 156) = 44.258$, $p < .001$, $r = .678$, $r^2 = .460$. Three variables were significant predictors in the regression: D, word hypernymy values, and CELEX content word frequency. Four variables were not significant predictors: word meaningfulness, word familiarity, LSA sentence to sentence, and word imagability. The latter variables were left out of the subsequent model; $t$-test information on these variables from the regression model as well as the amount of variance explained ($r^2$) is presented in Table 3. The results from the linear regression demonstrate that the combination of the three variables accounts for 46% of the variance in the human evaluations of lexical proficiency for the 160 essays examined in the training set (see Table 4 for additional information).
To further support the results from the multiple regression conducted on the training set, we used the B weights and the constant from the training set multiple regression analysis to estimate how the model would function on an independent data set (the 80 evaluated writing samples held back in the test set). The model produced an estimated value for each writing sample in the test set. We then conducted a Pearson correlation between the estimated score and the actual score. We used this correlation along with its $r^2$ to demonstrate the strength of the model on an independent data set. The model for the test set yielded $r = .649$, $r^2 = .421$. The results from the test set model demonstrate that the combination of the three variables accounted for 42% of the variance in the evaluation of the 80 writing samples comprising the test set.

### Total set model

To examine how the model from the training set predicted the variance in the human evaluations for the total corpus of writing samples, we used the B weights and the constant from the training set multiple regression analysis on the entire data set (the 240 evaluated writing samples that were contained in both the training and the test set). If the total set model is similar to the training and test set model, we can say with a high degree of confidence that it is a reliable model (Whitten and Frank, 2005). We followed the same methodology for the total set model as for the test set model. The model for the entire data set yielded $r = .666$, $r^2 = .444$. The results from the entire data set model demonstrate that the combination of the three variables accounts for 44% of the variance in the evaluation of the 240 writing samples that comprise the lexical proficiency corpus.

### Discussion

The results of this analysis demonstrate that three lexical indices taken from Coh-Metrix predict 44% of the variance of the human evaluations of lexical proficiency in the examined writing samples. The correlation value for the analysis ($r = .66$) demonstrates a large effect size indicating a strong relationship between the three lexical indices and the human evaluations of lexical proficiency (Cohen, 1992). The three variables, lexical diversity, word hypernymy, and word frequency are linked to both vocabulary size and depth of knowledge lexical features. No indices related to accessing core lexical items contributed significantly to our model of lexical proficiency, although many of these

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

<table>
<thead>
<tr>
<th>Entry</th>
<th>Variable added</th>
<th>Correlation</th>
<th>$r^2$</th>
<th>B</th>
<th>B (std)</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry 1</td>
<td>D</td>
<td>.584</td>
<td>.339</td>
<td>.022</td>
<td>.477</td>
<td>.003</td>
</tr>
<tr>
<td>Entry 2</td>
<td>Word hypernymy average</td>
<td>.641</td>
<td>.441</td>
<td>-1.130</td>
<td>-3.10</td>
<td>.218</td>
</tr>
<tr>
<td>Entry 3</td>
<td>CELEX content word frequency</td>
<td>.678</td>
<td>.460</td>
<td>-7.36</td>
<td>-2.45</td>
<td>.197</td>
</tr>
</tbody>
</table>

Notes: Estimated constant term is 4.701; $B$ is unstandardized Beta; $B$ (std) is standardized Beta; SE is standard error
indices demonstrated significant correlations with the human ratings. This finding has important implications for language testing and theories of lexical acquisition.

From a language assessment perspective, the results of this study present a model from which to evaluate the lexical proficiency of language learners. The model is supported through convergent validity, but more importantly, it does not suffer from a lack of construct validity because it is strongly linked to psychological theories of how the mind processes language. In addition, the model is also simple and practical enough that it can be used to inform pedagogical decisions and can be used to measure lexical development in language learners. The model for predicting lexical proficiency is presented below.

\[
\text{Predicted lexical proficiency} = 4.701 + (0.022 \times \text{lexical diversity: D value})
+ (-1.130 \times \text{average of word hypernymy value})
+ (-0.736 \times \text{content word frequency value})
\]

The above model relies on the use of Coh-Metrix. To apply the model, a user would input a text into the tool and collect the results for each of the indices used in the model. The user would then calculate the results from the model. The model will report a number between 1 and 5, which correspond to the lexical proficiency levels found in the survey instrument used by the human raters (see Appendix 1).

Unlike previous research into lexical proficiency, the results of this study allow us to extend them into a practical, computational model because the results do not rely on participant survey responses, which are time consuming, or test a relatively small number of words (Meara, 2005; Wesche and Paribakht, 1996; Schmitt and Meara, 1997). The natural evolution of survey approaches that attempt to measure multiple word features necessarily require researchers to develop more finely tuned tests with fewer words, thus making testing lexical proficiency over a range of words impractical. Such a conclusion led Meara and Wolter (2004) to decide that lexical proficiency could best be tested through a single measure of vocabulary size and not through depth of knowledge or access to core lexical items. However, unlike survey instruments meant to measure lexical proficiency, our approach allows us to examine relatively short texts produced by a learner using a variety of indices that provide adequate lexical coverage. In addition, these indices link to lexical features that are important to lexical proficiency and that go beyond vocabulary size.

From a theoretical perspective, this study provides evidence as to the strength of various lexical indices in predicting overall lexical proficiency. Perhaps the most robust finding of this study is that an index of lexical diversity, $D$, explains almost 34% of the variance in human judgments of written lexical proficiency. This signifies that human evaluations of lexical proficiency can best be predicted by the variety of words that a writer produces. Lexical diversity, as an indicator of vocabulary size, thus best explains proficient lexical use. In the writing samples analyzed, neither depth of word knowledge features nor access to core lexical items were as indicative as lexical diversity. Therefore, in our model of lexical proficiency the number of different words that the writer produces is the greatest contributor with more proficient lexicons characterized by more variety in word use.

The second most predictive lexical index for the sampled corpus was word hypernymy average, which explained just over 7% of the variance in our human ratings. Accordingly,
we argue that the level of specificity that a writer uses relates to judgments of lexical proficiency. More accurately, the more specific words a writer uses, the more likely the writer is to be judged as being less lexically proficient. Such a prediction is supported by recent research that demonstrates that L2 learners’ word use becomes less specific as time spent studying a language increases (Crossley et al., 2009). Thus, as writers develop hierarchical knowledge of word relations, allowing for less specific word use, they are evaluated as being more lexically proficient.

Lastly, this analysis has demonstrated that word frequency is also an important indicator of lexical proficiency. Our word frequency index accounted for about 5% of the variance in the human evaluations of lexical proficiency. Specifically, writing samples with less frequent words were evaluated as having greater lexical proficiency. This finding might be interpreted as supporting the importance of connectionist networks in lexical proficiency judgments. In such an interpretation, learners’ exposure to frequent words allows the words to be more easily accessible because of their repetition in language. The accessibility of frequent words, which develops over time, likely affords learners with greater processing resources to retrieve less frequent words. The use of less frequent words is thus associated with writers being more lexically proficient.

Of secondary interest are those lexical features that demonstrated significant correlations to the lexical proficiency ratings, but were not included in the final regression analysis. Although these indices were not predictive of lexical proficiency in our final model, they are still theoretically interesting. All of these indices were related to either depth of lexical knowledge features or access to core lexical items. They included the MRC psycholinguistic database indices (meaningfulness, familiarity, imagability, and concreteness) and our index of semantic co-referentiality taken from LSA. These indices demonstrated that writing samples scored as more lexically proficient contained less meaningful words, less familiar words, less imagable words, less concrete words, and less semantic co-referentiality. In general, these correlations support the notion that more lexically proficient writers use more sophisticated words that are more abstract, less familiar, less imagable, and have fewer semantic associations with other words.

Additionally, the indices that did not demonstrate significant correlations are of interest. For instance, our measure of polysemy did not demonstrate significant correlations with human judgments of lexical proficiency. This finding is likely the result of the index itself, which does not measure the frequency of the senses used or the number of senses for each word. Rather, the index reports on the number of senses that a word has in general, but not for the specific instance of the word as found in the context of the sample. Thus, we cannot exclude the importance of polysemy in models of lexical proficiency, but we can rule out the effectiveness of the WordNet polysemy measure in explaining human judgments of lexical proficiency. Also, polysemy features might be subsumed under the word frequency index that we used. It is generally accepted that more frequent words are also more polysemous (Crossley et al., 2010b). It is also important to note that traditional indices of word length showed no correlations with human judgments of lexical proficiency. We can therefore discount the notion that humans evaluate lexical proficiency based on word length.
Conclusion

This analysis has demonstrated that three lexical indices are predictive of human ratings of lexical proficiency. The indices, $D$, WordNet hypernymy, and CELEX content word frequency, explain 44% of the variance in human evaluations of lexical proficiency in the writing samples examined. These findings indicate strong relationships between the human evaluations and the lexical indices. The results of this study have important theoretical implications as well as the potential to transfer the resulting algorithm into an exploratory model from which to test lexical proficiency in language learners. Such explorations into lexical proficiency are critical because L2 lexical acquisition is enigmatic (Schmitt, 1998) and an overall theory of lexical acquisition is lacking (Meara, 1997). Additionally, the tools currently available from which to analyze lexical proficiency are time intensive and cannot examine a representative sample of words. The approaches described in this study address these problems by examining the efficacy of using computational indices that can examine larger word samples quickly and automatically.

This study also presents an approach that is rooted both in applied linguistics and psychology as called for by Meara (1997). We collected experimental data and analyzed it to produce a formal model that helps explain human judgments of lexical proficiency. We then used held back data to predict how well the data fits the predictions made by the formal model. We argue that the questions we ask and to which we provide answers are not only technically interesting, but also psychologically interesting in that they help to explain the interrelation of various lexical indices in explaining lexical proficiency.

There are, of course, elements of this current study that need to be further developed in order to improve the extendibility of the findings. First, our sample population of L2 learners was relatively small. With a smaller sample population, there should be less variance between the writing samples. We are optimistic that a larger sample population might actually increase the amount of variance explained by our selected computational indices. Second, many of the indices that we tested are not representative of the entire English lexicon. For example the MRC meaningfulness index we used contains human ratings for only 2627 words. While this might appear small, past analyses have demonstrated that the MRC indices provide coverage for about 80–85% of the word tokens found in speech (Salsbury et al., in press). We expect that those percentages would be lower for written language in which there is less lexical redundancy. Accordingly, many of the indices used in this study could benefit from additional lexical coverage. Third, some features of lexical proficiency were not included in this study because they have not been computationally implemented. Primary among these is an index of collocational proficiency. A study examining potential models of collocational proficiency has recently been completed (Crossley and Salsbury, in press). We are hopeful that such an index of collocational proficiency will prove an important predictor of lexical proficiency and strengthen the statistical findings reported in this paper. Last, this study only concentrated on written language. Further studies into the predictive validity of these indices for speech data are warranted.

While we present a model of lexical proficiency that might be considered rather simple, we have confidence in its predictive power. However, we also recognize that the model
we present does not explain all of the variance in the human ratings. This is likely the result of concentrating only on lexical indices related to vocabulary size, depth of knowledge, and access to core lexical items. We hypothesize that additional variance might be explained by indices that measure additional lexical features such as collocation proficiency and syntactic patterns. Additional variance might be explained by words’ pronounceability, orthography, and morphology (Nation, 2005). We argue, though, that this study forms a strong foundation for moving past exploratory considerations of a lexical organization system (Meara, 2002) and towards functional models of lexical proficiency. This is an important step in understanding what the term ‘lexical proficiency’ signifies and addressing methods of assessing it.

Acknowledgements

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We also give special thanks to our human raters: Brandi Williams, Jessica Mann, and Abigail Voller. We would also like to acknowledge the assistance of Dr. Phillip McCarthy in helping validate the scoring rubric used in this study.

References


**Appendix 1: Holistic rating form**

After reading each writing sample, assign a holistic score based on the rubric below. For the following evaluations you will need to use a grading scale between 1 (minimum) and 5 (maximum). The distance between each grade (e.g. 1–2, 3–4, 4–5) should be considered equal.

**Score of 5:** A sample in this category demonstrates clear and consistent mastery of the English lexicon, although it may have a few minor errors. A typical sample effectively uses appropriate conceptual categories (both concrete and abstract), demonstrates clear coherence between words, lexical-semantic connections, and is lexically diverse (enough to explain complex matters in detail). In general, the sample exhibits a skillful use of language, using a varied, accurate, and apt vocabulary with ease and fluency.

**Score of 4:** A sample in this category demonstrates reasonably consistent mastery of the English lexicon, although it will have occasional errors or lapses in lexical quality. The sample demonstrates the appropriate use of conceptual categories (both concrete and abstract), coherence between words, lexical-semantic connections, and lexical diversity that allows for the discussion of complex matters, but not always. Overall, the sample uses appropriate and precise vocabulary and appears fluent and accurate.

**Score of 3:** A sample in this category demonstrates adequate lexical mastery, although it will have lapses in quality. The sample demonstrates some appropriate uses of conceptual categories (including abstract concepts, but mostly concrete concepts), coherence between words, lexical-semantic connections, and lexical diversity. Overall, the sample uses generally appropriate and precise vocabulary, but demonstrates an inconsistent mastery.

**Score of 2:** A sample in this category demonstrates developing lexical mastery, but is marked by *one or more* weaknesses in conceptual categories, lexical-semantic connections, cohesion between words, and lexical diversity. However, discourse in the sample
is generally connected. By and large, the sample uses weak vocabulary or inappropriate word choice, lacks variety, and depends on concrete words. The lexical problems in the sample are serious enough that meaning is somewhat obscured.

**Score of 1:** A sample in this category demonstrates little lexical mastery, and is flawed by two or more weaknesses in conceptual categories, lexical-semantic connections, cohesion between words, and lexical diversity. The discourse in this category displays very little coherence and facility in the use of language. It also relies heavily on the use and repetition of memorized phrases. Overall, the sample demonstrates limited vocabulary, incorrect word choices, and exhibits frequent lexical problems so serious that meaning is often obscured.

Holistic score based on attached rubric (1–5): ___

**Appendix 2**