Lexical proficiency, as a cognitive construct, is poorly understood. However, lexical proficiency is an important element of language proficiency and fluency, especially for second language (L2) learners. For example, lexical errors are a common cause of L2 miscommunication (Ellis, 1995). Lexical proficiency is also an important attribute of L2 academic achievement (Daller, van Hout, & Treffers-Daller, 2003). Generally speaking, lexical proficiency comprises breadth of knowledge features (i.e., how many words a learner knows), depth of knowledge features (i.e., how well a learner knows a word), and access to core lexical items (i.e., how quickly words can be retrieved or processed; Meara, 2005). Understanding how these features interrelate and which features are important indicators of overall lexical proficiency can provide researchers and teachers with insights into language learning and language structure. Thus, this study investigates the potential for computational indices related to lexical features to predict human evaluations of lexical proficiency. Such an investigation provides us with the opportunity to better understand the construct of lexical proficiency and examine the capacity for computational indices to automatically assess lexical proficiency.

Recent investigations into lexical proficiency as both a learner-based and text-based construct have helped illuminate and model the lexical features
important in predicting lexical knowledge. From a learner-based perspective, Crossley and Salsbury (2010) demonstrated that word frequency, word familiarity, and word associations are the most predictive word features for noun production in beginning-level L2 learners. For early verb production, word frequency, word familiarity, word associations, and a word’s conceptual level were predictive of early L2 word production. These findings demonstrate that specific word features are predictive of lexical production and indicative of language proficiency for beginning learners. From a text-based perspective, Crossley, Salsbury, McNamara, and Jarvis (in press) used automated lexical indices to investigate variance in human judgments of lexical proficiency as found in L1 and L2 written texts. Crossley et al. found that human judgments of lexical proficiency were best predicted by a text’s lexical diversity, word frequency, and conceptual levels. This finding helps identify the individual lexical features important in explaining human judgments of lexical proficiency.

Such studies have contributed to defining lexical proficiency as a function of individual word properties. Additionally, studies such as these have important implications for lexical development, language acquisition, and pedagogical approaches. The studies help us understand the growth of lexicons in L2 learners and how these lexicons are an attribute of general language acquisition. The models extracted from these studies can be used to develop L2 academic assessments and to help classroom teachers develop lessons and materials that match the proficiency level of their students. However, empirical investigations into the construct of lexical proficiency such as these are scarce. Additional studies examining lexical proficiency are necessary. This is especially true for text-based assessments of spoken data. Spoken data, unlike written texts, are spontaneous and unmonitored and, thus, more likely reflect a speaker’s implicit lexical proficiency as compared to written or planned language production.

**METHODOLOGY**

The purpose of this study is to investigate the potential of automated lexical indices related to vocabulary size, depth of knowledge, and access to core lexical items to predict human ratings of lexical proficiency in spoken transcripts. We do so by analyzing a corpus of scored speech samples using lexical indices taken from the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004).

**Corpus Collection**

L2 speech samples were collected longitudinally from 29 participants at two different universities. The L2 participants ranged in age from 18
to 40 years and came from a variety of L1 backgrounds (Korean, Arabic, Mandarin, Spanish, French, Japanese, and Turkish). The students were enrolled either as undergraduates or in intensive language programs at the universities. The speech samples were transcribed from recorded conversations involving dyads of native speakers and nonnative speakers. The conversations were naturalistic and characterized by interactional discourse wherein ideas were shared freely and participants spoke on a variety of topics.

To ensure a range of lexical proficiency across the speech samples, the Test of English as a Foreign Language (TOEFL) or ACT ESL Compass scores were used to classify the L2 speakers’ samples into beginning, intermediate, and advanced categories. The classification of the L2 speakers, using the test maker’s suggested proficiency levels and descriptors, is as follows: L2 speakers who scored 400 or below on the TOEFL paper-based test (PBT), 32 or below on the TOEFL Internet-based test (iBT), or 126 or below on combined Compass ESL reading/grammar tests were classified as beginning level. L2 speakers who scored between 401 and 499 on the TOEFL PBT, 33 and 60 on the TOEFL iBT, or 127 and 162 on combined Compass ESL reading/grammar tests were classified as intermediate level. L2 speakers who scored 500 or above on the TOEFL PBT, 61 or above on the TOEFL iBT, or 163 or above on combined Compass ESL reading/grammar tests were classified as advanced level. In total, 180 L2 speech samples were collected from the participants to form the L2 corpus of speech data used in this study (60 samples for each proficiency level: beginning, intermediate, and advanced).

A comparison corpus of 60 native speech samples was selected from the Switchboard corpus (Godfrey & Holliman, 1993). The Switchboard corpus is a collection of about 2,400 telephone conversations taken from 543 speakers from all areas of the United States. The conversations are two-sided and involve a variety of topics. Like our L2 speech data, the speech samples in the Switchboard corpus are naturalistic. All speech samples extracted from the native speaker corpus were classified under the category native speaker.

In total, our speech sample corpus contained 60 speech samples from each L2 proficiency level as well as 60 speech samples from native speakers (N = 240). Two sets of speech samples were created. The first set contained only the utterances of the speaker in question (the L2 speaker or the native speaker). This set was used for the computational and statistical analysis. The second set contained the utterances of the speaker in question and those of the interlocutor. This set was used by the human raters to score the sample. The samples in the first set were controlled for length by randomly selecting a speech segment from each sample that was about 150 words. The speech samples were separated at the utterance level so that continuity was not lost. 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 was adapted from the American Council on the Teaching of Foreign Languages’ (ACTFL) proficiency guidelines for speaking and writing (ACTFL, 1999) and holistic proficiency rubrics produced by American College Testing (ACT) and the College Board. The survey was the same survey used in Crossley et al. (in press). The survey asked raters to evaluate lexical proficiency using a 5-point Likert scale. The survey defined high lexical proficiency as demonstrating clear and consistent mastery of the English lexicon such that word use was accurate and fluent and characterized by the appropriate use of conceptual categories, lexical coherence, lexical–semantic connections, and lexical diversity. The rubric defined low lexical proficiency as demonstrating little lexical mastery such that word use was flawed by two or more weaknesses in conceptual categories, lexical coherence, lexical–semantic connections, and lexical diversity that lead to serious lexical problems that obscured meaning (see Crossley et al., in press for the survey).

**Human Ratings**

To assess the 240 speech samples that make up our speech corpus, three native speakers of English were trained as expert raters using the lexical proficiency survey instrument. The raters were first trained on an initial selection of 20 speech samples taken from a training corpus not included in the speech corpus used in the study. The raters assigned each speech sample a lexical proficiency score between 1 (low) and 5 (high). After training, the raters scored the entire speech corpus. 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. After the training, the average correlation between the three raters for the entire corpus was $r = 0.808 (p < 0.001)$ with a weighted correlation of $r = 0.927$.

**Variable Selection**

All the lexical indices used in this study were provided by Coh-Metrix (Graesser et al., 2004). The lexical indices reported by Coh-
Metrix can be separated into vocabulary size measures, depth of knowledge measures, and access to lexical core measures. The measures related to vocabulary size include indices of lexical diversity. The measures related to depth of knowledge include indices related to lexical network models such as hypernymy, polysemy, semantic coreferentiality, word meaningfulness, and word frequency. Those measures related to accessing core lexical items include word concreteness, word familiarity, and word imagability. We also include indices related to word length for comparison purposes. This provided us with a total of 10 measures.

Measures

Many of the following measures report values for content words and for all words in the text (e.g., frequency, meaningfulness, imagability, concreteness, and familiarity). In these instances, we only selected the content word indices because of the tendency for speech to contain a greater incidence of function words resulting from disfluencies in speech.

Lexical Diversity

Lexical diversity indices generally measure the number of types (i.e., unique words occurring in the text) by tokens (i.e., all instances of words), forming an index that ranges from 0 to 1, where a higher number indicates greater diversity. Traditional indices of lexical diversity are highly correlated with text length and are not reliable across a corpus of texts where the token counts differ markedly. To address this problem, a wide range of more sophisticated approaches to measuring lexical diversity have been developed. Those reported by Coh-Metrix include MTLD and D (McCarthy & Jarvis, 2010).

Polysemy

Coh-Metrix measures word polysemy (the number of senses a word has) through the WordNet computational, lexical database (Fellbaum, 1998). Coh-Metrix reports the mean WordNet polysemy values for all content words in a text.

Hypernymy

Coh-Metrix also uses WordNet to report word hypernymy (i.e., word specificity). Coh-Metrix reports WordNet hypernymy values for all content words, nouns, and verbs on a normalized scale with 1 being
the highest hypernym value and all related hyponym values increasing after that. Thus, a lower value reflects an overall use of less specific words, while a higher value reflects an overall use of more specific words.

**Semantic co-referentiality**

Coh-Metrix measures semantic co-referentiality using Latent Semantic Analysis (LSA; Landauer, McNamara, Dennis, & Kintsch, 2007), which measures associations between words based on semantic similarity. Coh-Metrix reports on LSA values between adjacent sentences and paragraphs and between all sentences and paragraphs.

**Word frequency**

Word frequency indices measure how often particular words occur in the English language. The indices reported by Coh-Metrix are taken from CELEX (Baayen, Piepenbrock, & Gulikers, 1995), a 17.9 million-word corpus.

**Word concreteness**

Coh-Metrix calculates word concreteness using human word judgments taken from the MRC Psycholinguistic Database (Wilson, 1988). A word that refers to an object, material, or person generally receives a higher concreteness score than an abstract word.

**Word familiarity**

The MRC database also reports familiarity scores. Higher scores indicate greater familiarity. For example, the word *while* receives a mean familiarity score of only 5.43, whereas the more familiar word *eat* has a mean score of 6.71.

**Word imagability**

In addition, the MRC database reports imagability scores. A highly imagable word such as *cat* evokes images easily and is thus scored highly on the scale. A word such as *however* produces a mental image with difficulty and is thus scored lower on the scale.
**Word meaningfulness**

Coh-Metrix calculates word meaningfulness using the MRC Database. Words with high meaningfulness scores are highly associated with other words (e.g., *people*), whereas a low meaningfulness score indicates that the word is weakly associated with other words (e.g., *lathe*).

**Word length**

Word length has been a measure of word difficulty often employed in the past. Coh-Metrix reports two indices of word length: the average number of letters in a word and the average number of syllables in a word.

**ANALYSIS AND FINDINGS**

For our analysis, 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 speech sample. These variables were later used to predict the human scores in the training set using a linear regression model. Later, the speech samples 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.

To allow for a reliable interpretation of the multiple regression, we ensured that there were at least 20 times more cases (speech samples) than variables (the lexical indices). We used Pearson’s correlations to select a variable from each measure to be used in a multiple regression. Only those variables were selected that demonstrated significant correlations with the human ratings. With 160 ratings in the training set, this allowed us to choose 8 lexical variables out of the 10 selected measures. To check for multicollinearity, we conducted Pearson correlations on the selected variables. If the variables did not exhibit collinearity \((r < 0.70)\), they were then used in the multiple regression analysis.

**Pearson Correlations Training Set**

We selected the Coh-Metrix index from each measure that demonstrated the highest Pearson correlation when compared to the human ratings of the speech samples. The 10 selected variables and their measures along with their \(r\) values and \(p\) values are presented in Table 1,
sorted by the strength of the correlation. All measures produced at least one significant index.

Collinearity

Pearson correlations demonstrated that word concreteness was highly correlated (> 0.70) with the word imagability score ($N = 160$, $r = 0.930$, $p < 0.001$). Because the word concreteness value had a lower correlation with the human ratings as compared to the word imagability score, the word concreteness index was dropped from the multiple regression analysis. As a result, the eight indices that demonstrated the highest correlations with the human scores excluding concreteness were included in the regression analysis.

Multiple Regression Training Set

A linear regression analysis was conducted using the eight variables. These eight variables were regressed onto the raters’ evaluations for the 160 speech 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.

The linear regression yielded a significant model, $F(4, 155) = 61.831$, $p < 0.001$, $r = 0.784$, $r^2 = 0.615$. Four variables were significant predictors in the regression: $D$, word imagability, word familiarity, and word hypernymy. Four variables were not significant predictors: word frequency, word polysemy, LSA sentence to sentence, and word meaningfulness. The latter variables were left out of the subsequent model; $t$ test information on these variables from the regression model.

**TABLE 1**

Final Variables: Pearson Correlations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Measure</th>
<th>$r$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>Lexical diversity</td>
<td>0.679</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Word meaningfulness</td>
<td>MRC database</td>
<td>-0.512</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Word familiarity</td>
<td>MRC database</td>
<td>-0.466</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Word imagability</td>
<td>MRC database</td>
<td>-0.454</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Word concreteness</td>
<td>MRC database</td>
<td>-0.405</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Word hypernymy average</td>
<td>Hyponymy</td>
<td>-0.387</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CELEX content word frequency sentence</td>
<td>Word frequency</td>
<td>-0.291</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>LSA sentence to sentence adjacent</td>
<td>Semantic co-referentiality</td>
<td>-0.266</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Word polysemy average</td>
<td>Polysemy</td>
<td>0.244</td>
<td>&lt; 0.010</td>
</tr>
<tr>
<td>Average syllables per word</td>
<td>Word length</td>
<td>-0.179</td>
<td>&lt; 0.050</td>
</tr>
</tbody>
</table>
as well as the amount of variance explained ($r^2$) are presented in Table 2. The results from the linear regression demonstrate that the combination of the four variables accounts for 62% of the variance in the human evaluations of lexical proficiency for the 160 speech samples examined in the training set (see Table 3 for additional information).

**Test Set Model**

To further support the results from the multiple regression conducted on the training set, we used the B weights and the constant from the training set multiple regression analysis to estimate how the model would function on an independent data set (the 80 evaluated speech samples held back in the test set). The model produced an estimated value for each speech 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 $r = 0.772$, $r^2 = 0.595$. The results from the test set model demonstrate that the combination of the three variables accounted for 60% of the variance in the evaluation of the 80 speech samples comprising the test set.

**DISCUSSION AND IMPLICATIONS**

This analysis has demonstrated that four lexical indices are predictive of human judgments of lexical proficiency in speech samples. The indices measure breadth of knowledge features, depth of knowledge features, and access to core lexical items. Lexical diversity was the most predictive index and explained over 45% of the human ratings. Thus, the diversity of words in a sample best explains human judgments of lexical proficiency with high lexical proficiency samples containing a greater variety of words. Lexical diversity was followed by word

<table>
<thead>
<tr>
<th>Variable</th>
<th>$t$</th>
<th>$p$</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>9.004</td>
<td>&lt; 0.001</td>
<td>0.457</td>
</tr>
<tr>
<td>Word imagability</td>
<td>−2.680</td>
<td>&lt; 0.010</td>
<td>0.080</td>
</tr>
<tr>
<td>Word familiarity</td>
<td>−5.044</td>
<td>&lt; 0.001</td>
<td>0.052</td>
</tr>
<tr>
<td>Word hypernymy average</td>
<td>−3.284</td>
<td>&lt; 0.001</td>
<td>0.027</td>
</tr>
<tr>
<td>CELEX word frequency sentence</td>
<td>−0.672</td>
<td>&gt; 0.050</td>
<td>0</td>
</tr>
<tr>
<td>Word polysemy</td>
<td>1.621</td>
<td>&gt; 0.050</td>
<td>0</td>
</tr>
<tr>
<td>LSA sentence to sentence adjacent</td>
<td>−0.106</td>
<td>&gt; 0.050</td>
<td>0</td>
</tr>
<tr>
<td>Word meaningfulness</td>
<td>−0.550</td>
<td>&gt; 0.050</td>
<td>0</td>
</tr>
</tbody>
</table>

**TABLE 2**

Statistics ($t$ Values, $p$ Values, and Variance Explained) for Training Set Variables
imagability, which explained 8% of the human ratings. Speech samples judged as having greater lexical proficiency contained less imagable words. Word familiarity followed word imagability and explained 5% of the variance in the human scores. Samples evaluated as representing higher lexical proficiency contained more words that were less familiar. Last, word hypernymy explained 3% of the human scores of lexical proficiency and indicated that samples judged as more lexically proficient contained less specific words.

Overall, these findings portray greater lexical proficiency in speech data as the ability to use a wide range of words that evoke images less easily and are less familiar. The words are also less specific. Such a finding supports the notion that greater lexical proficiency is characterized by knowledge of more words, stronger lexical networks (i.e., hypernymy), and the production of words that are not easily retrievable (i.e., less imagable and familiar). This finding is in contrast to speech samples that exhibit lower lexical proficiency. These samples contain more lexical repetition and words that are more imagable, familiar, and specific. Thus, lower lexical proficiency is characterized by less variety of words. These words invoke features of core lexical items (i.e., imagability and familiarity), but are not strongly indicative of developed lexical networks. These findings differ from Crossley et al. (in press) in that judgments of lexical proficiency in spoken data, as compared to written texts, are better predicted by indices related to the accessibility of core lexical items. Such a finding likely relates to the contextual nature of spoken data, which may mitigate the need for highly imagable and familiar words.

The results of this study have strong practical implications. As we begin to better understand the construct of lexical proficiency, it affords us the opportunity to develop better tools for lexical instruction and assessment. Knowing how lexicons progress across learner levels can influence the development of learning material and teaching curriculum so that the expected lexical output of learners better matches their abilities. Statistical models of lexical proficiency also permit the development of lexical assessments from which to evaluate the lexical proficiency of learners.

### TABLE 3
Linear Regression Analysis to Predict Sample Ratings in Training Set

<table>
<thead>
<tr>
<th>Entry</th>
<th>Variable added</th>
<th>Correlation</th>
<th>$r^2$</th>
<th>$B$</th>
<th>$B_S$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$D$</td>
<td>0.676</td>
<td>0.457</td>
<td>0.032</td>
<td>0.496</td>
<td>0.004</td>
</tr>
<tr>
<td>2</td>
<td>Word imagability</td>
<td>0.732</td>
<td>0.536</td>
<td>$-0.006$</td>
<td>$-0.166$</td>
<td>0.002</td>
</tr>
<tr>
<td>3</td>
<td>Word familiarity</td>
<td>0.767</td>
<td>0.588</td>
<td>$-0.033$</td>
<td>$-0.273$</td>
<td>0.007</td>
</tr>
<tr>
<td>4</td>
<td>Hypernymy</td>
<td>0.784</td>
<td>0.615</td>
<td>$-0.688$</td>
<td>$-0.200$</td>
<td>0.209</td>
</tr>
</tbody>
</table>

Note: Estimated constant term is 24.69; $B$ is unstandardized Beta; $B_S$ is standardized Beta; SE is standard error.
proficiency of L2 learners. Such assessments, based on computational algorithms, can evaluate lexical proficiency quickly and automatically and inform student placement and assess their lexical development.

ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences (IES R305A080589). 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.

THE AUTHORS

Scott Crossley is an Assistant Professor at Georgia State University, where he teaches undergraduate and graduate courses in second language acquisition and general linguistics. His interests include computational linguistics, corpus linguistics, psycholinguistics, and second language acquisition. He has published articles in second language lexical acquisition, second language reading, second language writing, multidimensional analysis, discourse processing, speech act classification, and text linguistics.

Tom Salsbury, PhD, is on the faculty of the Department of Teaching and Learning at Washington State University. His research is in second language acquisition, and he publishes on vocabulary development among adults and children in content areas and also publishes in the area of pragmatics and language learning.

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 on a variety of topics, including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms.

Scott Jarvis, PhD, is an Associate Professor in the Department of Linguistics at Ohio University. His research has focused on methodological and theoretical challenges in the investigation of cross-linguistic influence, lexical diversity, and the measurement of language proficiency.

REFERENCES


