Applied Linguistics Advance Access published February 10, 2014

Applied Linguistics 2014: 1–22
doi:10.1093/applin/amt056

Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy

[1,]*SCOTT A. CROSSLEY, [2]TOM SALSBURY and [3]DANIELLE S. McNAMARA

[1]Department of Applied Linguistics/ESL, Georgia State University, [2]Department of Teaching and Learning, College of Education, Washington State University and [3]Department of Psychology, Arizona State University
*E-mail: sacrossley@gmail.com

This study analyzes lexical proficiency in oral and written texts produced by second language (L2) learners of English. The purpose of the study is to examine relationships between analytic scores of depth of lexical knowledge, breadth of lexical knowledge, and access to core lexical items and holistic scores of lexical proficiency. A corpus of 240 spoken texts and 240 written texts produced by beginning L2 learners, intermediate L2 learners, advanced L2 learners, and native speakers were scored for both analytic and holistic features of lexical proficiency by trained raters. Using a multiple regression analysis, the study found that collocation accuracy, lexical diversity, and word frequency are significant predictors of human evaluations of lexical proficiency. Collocation accuracy explained the greatest amount of variance in the holistic scores (84% in the written samples and 89% in the spoken samples). The authors discuss the importance of collocation accuracy in predicting human judgments of lexical proficiency.

INTRODUCTION

Lexical proficiency[1] has been described as *a mystery* (David 2008), *impossible to describe as a unified construct* (Henriksen 1999), *extremely intricate* (Boers *et al.* 2006), *poorly understood* (Crossley *et al.* 2011b), and *a complex phenomenon* (Zareva *et al.* 2005). While difficult to describe succinctly, the importance of lexical proficiency in language processing, linguistic awareness, and linguistic growth cannot be understated. Attaining lexical proficiency is one of the basic elements of language learning and development (David 2008), especially for second language (L2) learners for whom it is an important indicator of academic achievement (Daller *et al.* 2003). Lexical proficiency is also important in spoken communication where miscommunication frequently results from lexical errors (Ellis 1995) and in written communication where lexical errors are one of the most common problems that L2 writers face (Cumming 1990; Manchon *et al.* 2007).

Knowing the importance of lexical proficiency, many second language researchers have striven to develop valid computational methods with which to measure it (e.g. Malvern *et al.* 2004; Daller *et al*. 2007; McCarthy and Jarvis 2010; Crossley *et al.* 2011a, b). Such methods attempt to address the practical and theoretical questions about lexical proficiency common in second language acquisition and to help better define the construct. Despite these efforts, reliable methods for measuring lexical proficiency have been difficult to develop due largely to the number of lexical features inherent in vocabulary knowledge. Nonetheless, recent studies that investigated human judgments of lexical proficiency using computational methods have shown promise (Crossley *et al.* 2011a, b). The current study furthers these efforts by using human analytic judgments of lexical proficiency for common lexical features (e.g. human judgments of collocation accuracy, word frequency, word specificity) to predict holistic judgments of lexical proficiency. Such an analysis can provide a richer understanding of which elements of the lexicon are most predictive of lexical proficiency by tapping into human intuition to assess lexical proficiency as compared to relying on computational algorithms. Information gleaned from such an analysis can inform definitions of lexical proficiency and build on previous investigations that have examined this complex construct.

Lexical proficiency

There is a general consensus among many L2 researchers that lexical proficiency can be assessed using global trait models. These models primarily examine two dimensions that account for the majority of lexical knowledge (Henriksen 1999; Zareva *et al.* 2005). These dimensions are breadth of lexical knowledge (i.e. the size of a vocabulary) and depth of lexical knowledge (i.e. the manner and degree to which known words are organized; Meara 1996, 2005a; Read 1998). While these dimensions have been theoretically informative, not all elements of lexical proficiency can easily be slotted into these two categories, especially when orthographic, phonological, grammatical, and appropriateness components (Read 2000; Nation 2001) or differences between receptive and productive lexical knowledge are considered (Baba 2009). Also, comparisons between the two dimensions are difficult because breadth of lexical knowledge is characterized by knowledge of the lexicon, while depth of lexical knowledge applies to knowledge of individual words. Finally, the two categories exclude some lexical properties related to accessing core lexical items. Such properties include word concreteness, imageability, and familiarity and account for additional elements of words that allow for quicker lexical processing or retrieval (Meara 2005b; Crossley *et al.* 2011a, b).

It is important to note that these categories are theoretical-level constructs. Most researchers assume that at the observation level (i.e. at the level of language production), these categories are grouped into behavioral constructs such as lexical diversity and lexical sophistication (i.e. measures of theoretical-level constructs). Most lexical proficiency researchers then operationalize these behavioral constructs into lexical dimensions using computational algorithms (Bulté *et al.* 2008). Such algorithms assess behavioral constructs such as lexical diversity through type-token ratio counts or lexical sophistication through word frequency counts taken from large-scale corpora.

Unfortunately, there does not appear to be a concise agreement on which elements of word knowledge fit into which lexical dimensions. For example, some researchers argue that lexical diversity indices measure breadth of knowledge (Crossley *et al.* 2011a). This follows from the assumption that learners with a larger vocabulary produce more word types, and that vocabulary size is a measure of breadth of knowledge. However, in addition to measuring breadth of knowledge, lexical diversity can be considered an individual construct (i.e. a measure outside of breadth and depth of lexical knowledge; Baba 2009) as well as a measure of discourse cohesion (McCarthy and Jarvis 2010). As another example, word frequency is sometimes included within the dimension breadth of lexical knowledge under the assumption that learners who produce less frequent words know a greater number of words. At the same time, word frequency indices may also relate to the associative strength of words because word repetition in frequent words strengthens the connections between a word and its meaning (Ellis 2002), placing word frequency under depth of lexical knowledge.

Overall, all these features of lexical proficiency are important in affording a better understanding of how L2 learners process and produce language (Laufer and Nation 1995; Schmitt 1998; Qian 1999), and such features can provide researchers and teachers with important insights about how L2 learners acquire words and word meanings. While most L2 lexical research has focused on breadth of knowledge features, fewer studies have examined depth of knowledge features (Baba 2009) and even fewer have investigated access to core lexical items (Crossley *et al.* 2011a, b).

Previous studies of lexical proficiency

One difficulty in measuring lexical proficiency as a global trait is deciding how to measure proficiency. There are three common approaches to measuring lexical proficiency found in the literature: (i) assessing longitudinal changes in lexical features, (ii) investigating differences between proficiency levels, and (iii) examining human ratings of lexical proficiency.

Perhaps the most common approach to investigating lexical proficiency is examining how lexical production changes as a function of time spent studying English (i.e. longitudinal change). Schmitt (1998) collected interview data from a small group of English-as-a-second-language (ESL) participants over the course of a single academic year and found that the participants' word sense knowledge increased over the course of the study. David (2008), examining French-as-a-second-language learners over a 5-year period, found that lexical diversity increases as a result of learners proceeding from beginning to advanced stages of language learning. Bulte *et al.* (2008) also investigated lexical growth in French-as-a-second-language learners. They reported that the learners in their study demonstrated significant progression over time in terms of lexical diversity, lexical sophistication (i.e. frequency), and lexical productivity (content word production) over a 2-year period. Authors (2009) conducted a longitudinal study of six ESL learners and examined differences in their lexical diversity and word specificity. They found that over the course of a year, the learners produced a greater diversity of words that became less specific. In a similar study, Crossley *et al.* (2010a) examined the growth of sense relations in a longitudinal study of ESL learners. Like Schmitt (1998) they reported that L2 learners increased the number of senses attributed to individual words over the course of the study. Finally, Salsbury *et al.* (2011c) used a longitudinal approach to investigate the use of concrete words, imageable words, meaningful words, and familiar words (i.e. words with greater exposure) in ESL learners. They found that the learners produced less concrete, imageable, and meaningful words over the year-long study. They reported no differences for word familiarity.

Another common approach to investigating lexical proficiency is to examine differences in the lexical production of L2 learners from different proficiency levels (i.e. differences between beginning, intermediate, and advanced L2 learners). Among the first researchers to use this approach were Laufer and Nation (1995) who used lexical frequency profiles (LFP, a word frequency index) to assess differences between writing samples of L2 learners from three different proficiency levels. Laufer and Nation reported that LFP could be used to distinguish between the writers' proficiency levels and that learners at the advanced levels produced less frequent words than learners in lower proficiency levels. In a similar fashion, Zareva *et al.* (2005) used lexical indices related to quantity and quality of vocabulary to distinguish between intermediate and advanced L2 learners. They reported that vocabulary size, word frequency, and number of associations were the strongest indicators of proficiency level with L2 learners at the advanced levels demonstrating a larger vocabulary that included more infrequent words with more associations. More recently, Crossley *et al.* (2012a) investigated lexical differences in the writing samples of L2 learners from beginning, intermediate, and advanced levels. They reported that advanced learners produced less imageable words, more infrequent words, and used greater lexical diversity.

The last approach to assessing lexical proficiency is through the analysis of human ratings of lexical proficiency. Two recent studies by Crossley *et al.* (2011a, 2011b) best exemplify this approach. In both studies, the authors collected holistic judgments of lexical proficiency from expert raters. In Crossley *et al.* (2011a), expert ratings were collected for written samples and a statistical model was developed using lexical indices reported by the computational tool Coh-Metrix (Graesser *et al.* 2004) to predict these ratings. The model reported that human judgments of lexical proficiency were best predicted by a text's lexical diversity, word frequency, and specificity, with speakers rated as having greater lexical proficiency producing a greater amount of lexical diversity, less frequent words, and less specific words. In Crossley *et al.* (2011b), human judgments were collected for speech samples. A regression model based on lexical indices was then computed. This model demonstrated that speakers rated as having higher lexical proficiency by human raters produced greater lexical diversity, less imageable words, less familiar words and less specific words.

Overall, these studies demonstrate that lexical proficiency as measured through longitudinal growth, differences in proficiency level, and human ratings of lexical knowledge can be partially defined using lexical features (e.g. word sense knowledge, lexical diversity, word frequency, word concreteness, and word imageability). These studies have contributed to helping better define lexical proficiency as a function of individual word properties. These studies have also helped us to better understand lexical development in L2 learners. However, as stated above, lexical proficiency is difficult to define and reliable methods to assess it have proven difficult to develop. Thus, further examinations of lexical proficiency using different methods of analysis are needed to better understand the elements that underpin lexical proficiency and to help increase our general knowledge of the construct.

The purpose of this study is to investigate lexical proficiency by examining relationships between holistic scores of lexical proficiency and analytic scores of individual lexical features related to depth of lexical knowledge, breadth of lexical knowledge, and access to core lexical items. Such an approach will provide information about how human judgments of lexical elements in a text can be used to predict overall lexical proficiency. Such judgments may provide stronger links to behavioral constructs (e.g. collocational accuracy and sense frequency) than computational operationalizations of the same constructs. They will also provide alternative approaches to assessing lexical proficiency that, when combined with previous studies using computational indices, will provide additional criteria for defining lexical proficiency and identifying key elements of lexical proficiency.

METHODS

In this study, we conduct two analyses. The first analysis examines analytic and human scores derived from a corpus of written samples produced by L1 and L2 learners. The second analysis examines analytic and human scores derived from a corpus of speech samples produced by L1 and L2 learners. We examined both written and spoken speech samples to assess the potential for similarities and differences in how lexical proficiency is constructed in the two registers.

Corpus collection

We used the written corpus reported in Crossley *et al.* (2011a) and the spoken corpus reported in Crossley *et al.* (2011b). We thus focus on learner production with the understanding that the best way to investigate the lexicon is through its use, especially when informants are unaware that the lexicon is the focus of the assessment (Read 2000; Nation 2007). Both corpora used in this study contained L1 and L2 learner output. For the L2 learners, the samples were stratified by level (i.e. beginning, intermediate, and advanced learners) using Test of English as a Foreign Language or American College Test ESL Compass scores. An equal number of samples were collected from each level (beginning, intermediate, and advanced L2 learners). The L1 speakers in the samples were categorized as fluent speakers of the language. We sampled texts from these four levels to ensure a distribution of proficiency levels.

Writing corpus. The L2 writing samples from Crossley *et al.* (2011a) were collected from participants involved in a longitudinal study in an intensive language program at a large university in the United States. One hundred and eighty L2 writing samples were selected from a corpus of written data collected from 10 L2 learners over a 1-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. The selection process was based on creating a corpus stratified by language level.

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. 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). The majority of the samples were from Korean speakers ($n = 47$) followed by Arabic speakers ($n = 46$) and Japanese speakers ($n = 33$). Data were collected from a variety of L1s to create a corpus that was not L1 specific.

A matching L1 corpus was developed that consisted of 60 L1 freewrites collected from the Stream of Consciousness Data Set from the Pennebaker Archive Project (Newman *et al.* 2008). The stream of consciousness corpus includes thousands of free-writing samples collected from freshman psychology students. In a similar fashion to the L2 writing samples, the free-writing samples were unstructured and unprepared and written on a variety of topics. In total, the writing corpus contained 60 writing samples from each L2 proficiency level as well as 60 writing samples from the native speakers for a total of 240 sample texts. The samples were controlled for potential text length effects by randomly selecting a text segment from each sample that was about 140 words (depending on paragraph constraints). In this way, rater bias based on the length was controlled (i.e. as the texts were all of the same length, higher scores were not given to longer texts; Crossley *et al.* 2011a).

Spoken corpus. L2 speech samples from Crossley *et al.* (2011b) were collected longitudinally from 29 participants at two different universities ($n = 180$). 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 majority of the samples were from Arabic speakers ($n = 68$) followed by Korean speakers ($n = 43$) and Japanese speakers ($n = 23$). Like the written data, a variety of L1s were sampled to create a corpus that is not L1 specific. 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 thoughts and ideas were exchanged and participants spoke on a variety of topics. Like the written corpus, the selection process was based on creating a corpus stratified by language level.

A matching corpus of 60 native speech samples was selected from the Switchboard Corpus (Godfrey and 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. Unlike our L2 speech data, the Switchboard Corpus is not face-to-face conversations, but rather mediated by telephones. In total, the spoken corpus contained 60 speech samples from each L2 proficiency level (beginning, intermediate, and advanced levels) as well as 60 speech samples from native speakers ($N = 240$). Like the written samples, the spoken samples were controlled for text length effects by selecting samples of about 140 words depending on paragraph constraints. As reported in Crossley *et al.* (2011b), significant differences in holistic scores were not reported based on text length. For purposes of assessment, the speech samples contained the utterances of the speaker in question and those of the interlocutor.

Survey instrument

The survey instrument used in this study consisted of two sections. The first section prompted evaluations of analytical lexical features (e.g. *collocation accuracy, lexical diversity, word frequency*). The second section prompted an evaluation of holistic lexical proficiency based on lexical features theorized to be important in assessing lexical proficiency. The holistic rubric used in this study was 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 ACT and the College Board (for use in SAT writing evaluation). The holistic rubric is the same as that used and validated by Crossley *et al.* (2011a, 2011b). The analytic rubric is the same as used and validated by Authors (2013). The survey instrument is located in Appendix A (see online Supplementary material).

The analytic features in the survey instrument prompted raters to evaluate lexical features of theoretical interest in lexical proficiency research. The features relate to breadth of lexical knowledge, depth of lexical knowledge, and access to core lexical items. The features contained in the rubric were subcategorized as conceptual knowledge (*basic category use, word concreteness, word specificity*), lexical associations (*semantic co-referentiality, collocation accuracy, sense relations, sense frequency*), lexical frequency (*word frequency*), and lexical diversity (*type/token ratio*). These features and their relationship to lexical proficiency are discussed below.

Basic category words. The most prototypical lexical item within a hierarchy of superordinate and subordinate words is called a basic category word. For example, the word *fish* is a basic category word in the hierarchical scale represented by *animal* (a superordinate term) and *salmon* (a subordinate term). Basic category words contain a large number of cues or features to distinguish them from other words, and they are the words most often used in a hierarchical scale to discuss a concept. L2 learners tend to produce more basic category words in early stages of acquisition because these words are more frequent in the input and because they express more general meaning (Levenston and Blum 1977).

Word specificity. A related concept to basic category words is word specificity. As already mentioned, a basic category word such as *fish* contains the most cues or features to distinguish it from other terms at a similar level (*mammal, reptile, bird*). A more specific word (subordinate word) such as *salmon* has lower cue validity because many of the cues or features that distinguish *salmon* from other members of the category (*trout, pike, flounder*) are not shared across the category. Less specific words (superordinate words) related hierarchically to *fish* (such as *animal* or *entity*) have low cue validity because the available cues to distinguish words at this level are not strongly discriminatory (Rosch *et al.* 1976). Thus, basic category words are likely the point from which words deviate toward lower cue validity by either becoming more specific or less specific. Crossley *et al.* (2009) demonstrated that L2 learners move toward the production of less specific words (superordinate words) as a function of increasing linguistic proficiency.

Word concreteness. Words vary with regard to their level of concreteness. Words that are highly concrete refer to here-and-now concepts, ideas, and things (Paivio *et al.* 1968; Toglia and Battig 1978; Gilhooly and Logie 1980). Such words have advantages in tasks involving recall, word recognition, lexical decision, pronunciation, and comprehension (Paivio 1991; Gee *et al.* 1999). A word's level of concreteness also impacts its learnability. For example, studies have demonstrated that concrete words are learned earlier by L2 learners (Crossley *et al.* 2009; Salsbury *et al.* 2011). In addition, concrete words are learned more easily than abstract words (Ellis and Beaton 1993).

Semantic co-referentiality. Semantic co-referentiality refers to how words are semantically related beyond the morphological level. One example of semantic co-referentiality is synonyms. The words *dog* and *pooch* may serve similar functions in discourse because they are semantically similar; however, they are not related morphologically. The concept of *cat* may also include the words *tail, fur, claw,* and *whiskers.* These words are all unrelated morphologically, but they are connected semantically. Semantic similarity goes beyond conceptual associations as illustrated in the word pair *cat* and *mouse*. These two words are more closely linked semantically than *dog* and *mouse*. Crossley *et al.* (2010b) demonstrated that L2 learners' utterances develop stronger semantic links over time (i.e. L2 learners use a greater number of semantically related words).

Collocational accuracy. Collocations or multi-word lexical units refer to combinations of words that are expected and acceptable to speakers of a language. Such units may be adjacent lexical items or items that co-occur within a span of three or more words. Multi-word units are important lexical items acquired by L2 learners (Wray 2002, 2008) largely because of their role as indicators of communicative competence (Moon 1992; Lennon 1996). Multi-word units also express naturalness in language production in that listeners and readers are primed to expect certain words following exposure to another word. When listeners and readers hear an expected word, the language is more natural. When they do not hear an expected word, the language is less natural (Hoey 2005). Thus, our definition of collocational accuracy refers to how accurate learners are in producing acceptable and expected multi-word units. Multi-word units contain both lexical and syntactic information, and the use of multi-word units assist L2 learners to appear more proficient (Boers *et al.* 2006). Recent studies have found that competence with multi-word units facilitates effective and fluent communication (Nesselhauf 2003) and that multi-word unit accuracy develops with time spent studying English (Crossley and Salsbury 2011).

Sense relations. Sense relations are captured through word polysemy, which refers to the number of unique or related senses of a given word form. Interestingly, more frequent words generally have more senses (such as the word *get*) and are therefore more ambiguous (Davies and Widdowson 1974). Research demonstrates that word sense knowledge increases as L2 learners gain proficiency (Schmitt 1998) and that, as language proficiency increases, L2 learners produce words that are more polysemous as well as produce more senses for individual words (Crossley *et al.* 2010a).

Sense frequency. Some senses of words are more frequent than other senses. The noun *book* is a more frequent sense of this word form than the verb *book* for making a reservation or *book* for recording charges against a person in a police blotter. Thus, research on sense frequency in SLA distinguishes ambiguity as seen in multiple senses from the sense of the word that is produced. The frequency of the word sense that a learner produces may indicate the lexical competence of the speaker or writer. For this reason sense frequency is important to researchers. Past research has demonstrated that L2 learners produce less frequent word senses as a product of increasing linguistic proficiency (Schmitt 1998; Crossley *et al.* 2010a).

Word frequency. Word frequency refers to the frequency of words in a text regardless of how many senses the words possess. Structural regularities based on word frequency effects in language have important facilitative effects for lexical acquisition (Ellis 2002). Researchers in the field have long known that the highest frequency words account for the majority of linguistic tokens in any language sample and are relatively few in number. In contrast a vast majority of words occur very infrequently in a given text. This results in a Zipfian distribution (Zipf 1935). The Zipfian distributional bias optimizes language acquisition by providing high-frequency exemplars from which to learn linguistic constructions (Ellis and Collins 2009). The production of infrequent words is an important indicator of lexical knowledge, with more proficient lexicons characterized by the use of less frequent words (Daller *et al.* 2003; Crossley *et al.* 2011a, b).

Lexical diversity. Lexical diversity is a text-internal measure of lexical richness related to breadth of knowledge. It is a measure of the variety of words in a text using a modified type-token analysis. More proficient writers produce a greater variety of words (as measured through curve-fitting formulae; Jarvis 2002) as do more proficient speakers (Higgins *et al.* 2011). The variety of words produced, as measured by *D* (Malvern *et al.* 2004), is also a strong predictor of human judgments of overall lexical proficiency (Crossley *et al.* 2011a, b).

Human ratings

To assess the lexical features found in the 240 writing samples and the 240 speaking samples that comprise our written corpus, three native speakers of English were trained as expert raters. The 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 analytic feature a score between 1 (minimum) and 6 (maximum) and also assigned each writing sample a holistic score between 1 (minimum) and 5 (maximum). This process was then repeated with speech samples. The average score of the raters was then calculated for each sample, leading to 15 possible holistic scores between 1–6 including 1.33 and 1.66 and 12 analytic scores between 1–5 including 1.33 and 1.66.

To assess test reliability, we conducted a Cronbach's alpha for the 10 items in the survey instrument. The test reliability for the spoken data reported $\alpha = .732$ and the test reliability for the written data reported $\alpha = .724$, indicating acceptable reliability (Kline 1999). To assess inter-rater reliability (IRR), we report intra-class correlations between raters using Cronbach's alpha (see Table 1). For both corpora, the raters had the strongest agreement on collocation use and the lowest agreement on the use of basic category words.

Statistical analysis

We used training sets to develop models of lexical proficiency using the analytic scores as independent variables (i.e. predictors) and the holistic scores as the dependent variables. We only selected the analytic scores that demonstrated high IRR among the raters (i.e. those scores that had an $\alpha \geq .700$; Multon 2010). We first divided the data into training and test sets to assess the performance of the model on an independent corpus that was not used in the initial analysis. We observed the parameters set by Whitten and Frank (2005) by dividing the corpora based on a 67/33 split, leading to training sets of $n = 180$ and test sets of $n = 60$. We used the training sets to identify

Table 1: Intra-class correlations between human raters

Item	α Written corpus	α Spoken corpus
Basic category use	0.450	0.395
Word specificity	0.530	0.499
Word concreteness	0.771	0.807
Semantic co-referentiality	0.571	0.539
Collocation accuracy	0.919	0.905
Sense relations	0.599	0.426
Sense frequency	0.652	0.664
Word frequency	0.738	0.779
Lexical diversity	0.821	0.773
Holistic score	0.912	0.920

which of the analytic features best correlated with the holistic scores assigned to each sample. We conducted both Pearson product moment correlations and Spearman rank correlations. We conducted both correlational tests to control for the possibility that although the raters were trained to use interval scales, they may have used ordinal scales (i.e. they may not have treated the distance between the scores as equal). The variables that demonstrated significant correlations were then used to predict the human scores in the training set using a linear regression model. To calculate the predictive strength of the variables in an independent corpus, the remaining samples in the test set were analyzed using the regression model from the training set (Whitten and Frank 2005).

To avoid overfitting of the model and allow for a clear interpretation of the variables' individual contribution, we ensured that there were at least 20 times more cases (samples) than variables (the analytic scores). Such an approach affords a more reliable interpretation of the multiple regression results (Barcikowski and Stevens 1975; Field 2005). We also assessed for multicollinearity between the variables to ensure they were not measuring similar features. To assess multicollinearity between variables, we conducted additional correlations and checked both variance inflation factors (VIF) and tolerance levels. We set cutoffs for multicollinearity at $r < .90$ for the correlations (Tabachnick and Fidell 2001) and VIF values of around 1 and tolerance levels at the .2 threshold (Field 2005).

RESULTS

Written corpus

Pearson correlations training set. The correlations (both Pearson and Spearman) between the four analytic scores (collocational accuracy, lexical diversity, word

Table 2: Correlations: analytic scores to holistic scores for written corpus (training set)

Index	r	ρ
Collocation accuracy	0.914*	0.858*
Lexical diversity	0.703*	0.624*
Word frequency	−0.609*	−0.525*
Word concreteness	−0.102	−0.069

*$p < .001$.

Table 3: Correlations for multicollinearity: written data

Analytic feature	Collocational accuracy	Word frequency	Lexical diversity
Word concreteness	−0.027	0.197	−0.058
Collocational accuracy		−0.523	0.633
Word frequency			−0.637

frequency, and word concreteness) that demonstrated high IRR in the written corpus and the holistic scores of lexical proficiency are presented in Table 2. All analytical scores except *word concreteness* yielded significant correlations with the holistic scores. The strongest correlations were reported for *collocation accuracy* and *lexical diversity*.

Collinearity. According to the Pearson correlations, no variables demonstrated multicollinearity (see Table 3 for correlations). Thus, the three significant variables were included in the subsequent regression analysis.

Multiple regression training set. Using these three variables (collocational accuracy, lexical diversity, and word frequency), we conducted a linear regression analysis by regressing the analytic scores onto the holistic scores for the 160 writing samples in the training set. The linear regression using the three variables yielded a significant model, $F(3,\ 162) = 367.617$, $p < .001$, $r = .934$, $r^2 = .870$. All three variables were significant predictors in the regression. The combination of the three variables accounts for 87% of the variance in the holistic evaluations of lexical proficiency for the 160 essays examined in the training set (see Table 4 for additional information).

Test set model. To provide support for the results from the multiple regression model resulting from the training set data, we used the B weights and the constant from the analysis to examine how predictive the model would be on an independent data set (the 80 writing samples held back in the test

Table 4: Regression analysis: analytic scores predicting holistic scores for written corpus (training set)

Variable	r	r^2	ß	B	SE
Collocation accuracy	0.918	0.842	0.703	0.784	0.032
Lexical diversity	0.931	0.867	0.186	0.151	0.047
Word frequency	0.934	0.872	−0.111	−0.092	0.043

Constant = 0.148.

Table 5: Correlations: analytic scores to holistic scores for spoken corpus (training set)

Index	r	ρ
Collocation accuracy	0.944**	0.915**
Lexical diversity	0.784**	0.776**
Word frequency	−0.619**	−0.626**
Word concreteness	−0.332*	−0.338**

*$p < .050$, **$p < .001$.

set). We used the model to produce an estimated value for each sample in the test set. We then conducted a Pearson correlation between the estimated score reported by the model and the actual score reported by the human raters. This correlation along with its r^2 can be used to demonstrate the strength of the model on the independent data set. The model for the test set yielded $r = .916$, $r^2 = .839$, demonstrating that the combination of the four variables accounted for 84% of the variance in the human scores for the evaluation of the 80 writing samples comprising the test set.

Spoken corpus

Pearson correlations training set. The correlations (both Pearson and Spearman) between the four analytic scores (collocational accuracy, lexical diversity, word frequency, and word concreteness) that demonstrated high IRR in the spoken corpus and the holistic scores of lexical proficiency are presented in Table 5. All analytical scores yielded significant correlations with the holistic scores. The strongest correlations were reported for *collocation accuracy and lexical diversity*.

Collinearity. According to the Pearson correlations, no variables demonstrated multicollinearity (see Table 6 for correlations). Thus, the four significant variables were included in the subsequent regression analysis.

Multiple regression training set. The linear regression using the four variables (collocational accuracy, lexical diversity, word frequency, and word

Table 6: Correlations for multicollinearity: spoken data

Analytic feature	Collocational accuracy	Word frequency	Lexical diversity
Word concreteness	−0.338	0.207	−0.116
Collocational accuracy		−0.577	0.768
Word frequency			−0.524

Table 7: Regression analysis: analytic scores predicting holistic scores for written corpus (training set)

Variable	*r*	r^2	ß	B	SE
Collocation accuracy	0.942	0.886	0.664	0.738	0.034
Lexical diversity	0.951	0.905	0.266	0.191	0.049
Word frequency	0.954	0.909	–0.104	−0.079	0.039
Word concreteness	0.955	0.912	−0.074	−0.054	0.034

Constant = 0.371.

concreteness) yielded a significant model, $F(4, 161) = 417.268$, $p < .001$, $r = .955$, $r^2 = .910$. All four variables were significant predictors in the regression. The combination of the four variables accounted for 91% of the variance in the holistic evaluations of lexical proficiency for the 160 essays examined in the training set (see Table 7 for additional information).

Test set model. Using the model from the training set on the samples in the test set yielded $r = .948$, $r^2 = .899$, demonstrating that the combination of the four variables accounted for 90% of the variance in the human scores for the evaluation of the 80 speaking samples comprising the test set.

DISCUSSION

This study has demonstrated that analytic judgments of *collocation accuracy*, *lexical diversity*, and *word frequency* are highly predictive of holistic judgments of lexical proficiency for both written and spoken samples. In addition, analytic judgments of *word concreteness* are predictive of lexical proficiency in spoken samples. These findings indicate that depth of knowledge features related to collocation accuracy are the strongest predictors of lexical proficiency. Other significant predictors such as *lexical diversity* and *word frequency* are more strongly related to breadth of lexical knowledge, while *word concreteness* is related to access to core lexical items.

The analytic feature that explained the greatest amount of the variance in the holistic judgments of lexical proficiency was *collocation accuracy* (i.e. the words in the sample collocate accurately together). This feature explained 84% of the holistic scores in the written samples and 89% of the variance in the spoken samples. Such findings indicate the primacy that multi-word units have in explaining lexical proficiency. As Boers *et al.* (2006) reported, it appears that the accurate use of collocations indicates to expert raters that both the writer and the speaker have greater proficiency in lexical use and produce fewer lexical errors. Such a finding likely rests in the notion that multi-word units contain both lexical and syntactic components (i.e. syntagmatic and paradigmatic information) and that multi-word units are key indicators of fluent communication (Nesselhauf 2003) and communicative competence (Moon 1992; Lennon 1996).

Collocations are also likely stored as chunks in speakers' memory and thus are easier to retrieve and likely lead to more fluent language production, especially under real-time conditions such as speaking (Skehan 1998). These stored chunks may also lead to fewer lexical errors on the part of L2 learners (Boers *et al.* 2006), allowing language use to appear more natural and proficient (Hoey 2005). Previous models of lexical proficiency (e.g. Crossley *et al.* 2011a, b) did not include indices of collocation accuracy because computational operationalizations of this behavioral construct were unavailable. We hypothesize that if such indices could be developed and validated, they would help to explain a greater variance in human judgments when combined with other automated lexical indices related to lexical diversity, frequency, and specificity, providing us with a greater understanding of the role collocation accuracy plays in lexical proficiency and development.

The next most informative analytic feature in our models of lexical proficiency was *lexical diversity*, which explained about 3% of the variance in the written samples and about 2% of the variance in spoken samples. The *lexical diversity* scores correlated positively with the holistic scores indicating that writers and speakers that produce a greater variety of words are rated as having higher lexical proficiency. Such a finding supports the notion that L2 learners and L1 speakers that can produce a greater number of words (i.e. that have a greater breadth of lexical knowledge) are rated as more lexically proficient.

After *lexical diversity*, both the written and spoken models included *word frequency* as an indicator of lexical proficiency. For both written and spoken samples, *word frequency* explained about 1% of the variance. In both cases, writers and speakers who produced less frequent words were judged to have greater lexical proficiency. Since word frequency can be an indicator of breadth of knowledge, the findings may provide additional evidence that writers and speakers that have larger vocabularies are judged to be more lexically proficient.[2]

Finally, the model of spoken lexical proficiency included an index of word concreteness, which explained about 1% of the variance. This index was negatively correlated with lexical proficiency, indicating that speakers who produced less concrete words were judged to be more proficient. Less concrete

Table 8: Comparison between low- and high-scored writing sample: analytic scores

N	Level	Holistic score	Collocation accuracy	Lexical diversity	Word frequency
226	Native speaker	5.000	5.667	5.000	5.333
75	Beginning	1.333	2.667	2.333	3.666

N = number of participant.

words are less likely to be core lexical items and thus more difficult to access. Thus, the use of more difficult words indicates greater lexical proficiency.

To illustrate these differences, we present samples written by a native speaker of English and a beginning-level L2 learner. The samples differ in the expected patterns for the human rating assigned for collocation accuracy, lexical diversity, and word frequency (see Table 8 for human ratings for both texts).

Native speaker writing sample

> I want to have a smoke. I've been smoking for years and I just decided to quit about two hours ago. severe huh. "yeah me and Michele just fell out of touch about two hours ago" that's from Romy and Michele's high school reunion. good movie. it used to be the BEST movie, I'd watch it all the time with my best friend Niki. I use the term loosely, she was a great friend but moved away and never spoke to me again. stood me up a few times, no big deal. she's flakey. I don't consider myself flakey, I believe that I am very loyal and true to my word. for the most part, we all make mistakes, don't we? of course. I have the weirdest taste in my mouth, it's called FOOD, usually it's not this strong because I've got the Marlboro flavor burning my mouth up.

Beginning-level L2 writing sample

> My favorite country: UAE is a very nice country in the Gulf. In the UAE there is seven cities and the Capital of them is Abu Dhabi this where I live. Abu Dhabi is really the beautiful cities of this seven cities. There is the President where is live. The people in UAE is friend and the like to help anybody doesn't know about our country. The best thing I like there is all this seven cities is safety and anybody visit it would like to visit it more than one time.

The samples differ in terms of judgments of collocation accuracy. In the absence of automated indices to assess collocation accuracy, we conduct a limited qualitative analysis. As can be seen, the native speaker sample makes greater use of phrasal verbs such as *fell out*, *used to be*, *stood up*, and *burn up*. The native speaker also uses many common multi-word units such as *out of touch*,

Table 9: Comparison between low- and high-scored writing sample: coh-Metrix scores

N	Level	Holistic score	Lexical diversity: MTLD	Word frequency by sentence: CELEX
226	Native speaker	5.000	84.000	2.487
75	Beginning	1.333	26.554	3.026

N = number of participant.

high school reunion, all the time, best friend, moved away, no big deal, true to my word, for the most part, make mistakes, of course, and *weirdest taste*. Such collocations are not as evident in the L2 writing sample. There appears to be no use of phrasal verbs on the part of the L2 learner and few, if any, common multi-word units.

Unlike collocation accuracy, lexical diversity and word frequency can be measured automatically to some degree. These indices, as calculated by the computational tool Coh-Metrix, are reported in Table 7. For lexical diversity, we selected the Measure of Text and Lexical Diversity (MTLD: McCarthy and Jarvis 2010). For word frequency, we selected an index that measures the frequency of words in the Communitatis Europeae Lex corpus for each sentence in the text. The values reported by these indices (see Table 9) support an intuitive reading of the samples as found in the lexical diversity of the samples (i.e. that the native speaker uses a greater variety of words) and the word frequency of the words in the sample (i.e. that the native speaker uses less frequent words).

The models reported in this study support the notion that depth of knowledge features are the most important trait-based dimension in judgments of lexical proficiency (especially collocation accuracy). In both of our models, depth of knowledge features explained around 84–87% of the variance in lexical proficiency evaluations (if *word frequency* is included as a depth of knowledge measure). Breadth of knowledge features explained about 3% in the same models (around 4% if *word frequency* is included as a breadth of knowledge feature). Features that assessed access to core lexical items were not predictive in the written regression model reported, but *concreteness* was predictive of spoken lexical proficiency explaining about 1% of the variance in the human ratings. These findings indicate that it is not exactly the size of the vocabulary that counts, but the manner and degree to which words are organized in relation to each other when assessing lexical proficiency. Such a finding may indicate the importance of context in defining lexical proficiency. This finding, exemplified by the importance of our collocational accuracy feature, holds for both our written and spoken data even though the data sets are not meant to be comparable. The findings between the two data sets are commensurate indicating overlap in the features that best define lexical proficiency regardless of how the language is delivered.

It should also be noted that human raters might not be able to accurately come to agreement on many of the intrinsic features that theoretically help to define lexical proficiency. For the data reported in this study, five of the nine lexical features showed less than acceptable IRR among raters. These features included *basic category use*, *word specificity*, *semantic co-referentiality*, *sense relations*, and *sense frequency*. While agreement on these features was low (i.e. $\alpha < .70$), we, like Boers *et al.* (2006), are very much aware that measuring proficiency is an extremely difficult endeavor and that the range of properties inherent in a word render obtaining reliable scores difficult. However, it may be the case that computational methods are better suited than human raters for assessing some elements of lexical proficiency in some instances. For instance, Latent Semantic Analysis (Landauer *et al.* 2007) has proven a reliable measure of semantic co-referentiality in both written and spoken texts. In addition, indices derived from WordNet (Fellbaum 1998) have been shown to provide reliable measurements of word specificity and sense relations (Crossley *et al.* 2009, 2010a). In contrast, computational methods for assessing collocational accuracy in text are rare and the amount of variance such indices explain in human ratings is relatively low (e.g. Crossley *et al.* 2012b) compared with the strength reported for human ratings in this study. Thus, it is possible that the best approach to assessing lexical proficiency would mix both human and computer ratings.

CONCLUSION

Overall, the findings from this study help to better define the construct of lexical proficiency for naturalistic spoken and written texts.[3] Unlike past studies, this study focuses specifically on analytic features of lexical proficiency and how these features can be used to predict holistic ratings of lexical proficiency. The models of lexical proficiency reported in this study strongly support the notion that collocation accuracy on the part of both writers and speakers is the strongest predictor of lexical proficiency. Unfortunately, collocation accuracy is also one of the more difficult behavioral constructs to automatically measure at an operational level. Such difficulty arises from the sheer number of collocations available in a language and the semantic, grammatical, and contextual knowledge contained within a collocation that may or may not be explicit, thus making assessment difficult. As a result, automatically assessing lexical proficiency may be difficult until such time that collocation accuracy can be measured accurately. Such a finding indicates the importance of using human judgments to understand lexical proficiency. Such judgments provide us with access to lexical features that are more contextual in manner and assess the accurate and naturalistic use of words within a text. Currently, such measures are impossible to obtain computationally.

However, human judgments are not without their own limitations. As this study has demonstrated, expert human raters cannot always provide reliable ratings on a variety of lexical features. In addition, human raters require access to context (at least in the case of spoken data) to provide reasonable assessments of proficiency. Such context, while natural, provides interlocutor data, which may indicate lexical recycling and priming, both of which may influence human ratings. Regardless of the potential limitations of human ratings, human ratings seem to provide important assessments of lexical proficiency that are not available through other means. These assessments provide us with a fuller picture of the elements that help define lexical proficiency and the importance of individual lexical features in developing models of lexical proficiency.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

ACKNOWLEDGEMENTS

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. We also give special thanks to our human raters: Brandi Williams, Jessica Mann, and Abigail Voller.

NOTES

1 We use the term lexical proficiency as a blanket term that includes lexical knowledge and lexical competence.

2 Word frequency indices may also relate to depth of knowledge because more frequent words have a greater number of associations and less frequent words have fewer associations (Ellis 2002).

3 The results may change if the genre of the samples changed.

REFERENCES

Baba, K. 2009. 'Aspects of lexical proficiency in writing summaries in a foreign language,' *Journal of Second Language Writing* 18: 191–208.

Barcikowski, R. and **J. P. Stevens.** 1975. 'A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations,' *Multivariate Behavioral Research* 10: 353–64.

Boers, F., J. Eyckmans, J. Kappel, H. Stengers, and **M. Demecheleer.** 2006. 'Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test,' *Language Teaching Research* 10/3: 245–61.

Bulte, B., A. Housen, M. Pierrard, and **S. van Daele.** 2008. 'Investigating lexical proficiency development over time: The case of Dutch-speaking learners of French in Brussels,' *French Language Studies* 18: 277–98.

Crossley, S. A., T. Salsbury, and **D. S. McNamara.** 2009. 'Measuring second language lexical growth using hypernymic relationships,' *Language Learning* 59: 307–34.

Crossley, S. A., T. Salsbury, and **D. S. McNamara.** 2010a. 'The development of polysemy and frequency use in English second language speakers,' *Language Learning* 60: 573–605.

Crossley, S. A., T. Salsbury, and **D. S. McNamara.** 2010b. 'The development of semantic relations in second language speakers: a case for Latent Semantic Analysis,' *Vigo International Journal of Applied Linguistics* 7: 55–74.

Crossley, S. A., T. Salsbury, D. S. McNamara, and **S. Jarvis.** 2011a. 'What is lexical proficiency? Some answers from computational models of speech data,' *TESOL Quarterly* 45: 182–93.

Crossley, S. A., T. Salsbury, D. S. McNamara, and **S. Jarvis.** 2011b. 'Predicting lexical proficiency in language learners using computational indices,' *Language Testing* 28: 561–80.

Crossley, S. A. and **T. Salsbury.** 2011. 'The development of lexical bundle accuracy and production in English second language speakers,' *International Review of Applied Linguistics in Language Teaching* 49: 1–26.

Crossley, S. A., T. Salsbury, and **D. S. McNamara.** 2012a. 'Predicting the proficiency level of language learners using lexical indices,' *Language Testing* 29: 240–60.

Crossley, S. A., Z. Cai, and **D. S. McNamara.** 2012b. 'Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality' in P. M. McCarthy and G. M. Youngblood (eds): *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. The AAAI Press, pp. 214–19.

Cumming, A. 1990. 'Metalinguistic and ideational thinking in second language composing,' *Written Communication* 7: 482–511.

Daller, H., R. van Hout, and **J. Treffers-Daller.** 2003. 'Lexical richness in the spontaneous speech of bilinguals,' *Applied Linguistics* 24/2: 197–222.

Daller, H., J. Milton, and **J. Treffers-Daller.** 2007. *Modelling and Assessing Vocabulary Knowledge*. Cambridge University Press.

David, A. 2008. 'A developmental perspective on productive lexical knowledge in L2 oral interviews,' *French Language Studies* 18: 315–31.

Davies, A. and **H. Widdowson.** 1974. 'Reading and writing' in J. Allen and S. Corder (eds): *Techniques in Applied Linguistics*. Oxford University Press, pp. 154–201.

Ellis, N. 2002. 'Frequency effects in language processing,' *Studies in Second Language Acquisition* 24/2: 143–88.

Ellis, N. and **A. Beaton.** 1993. 'Psycholinguistic determinants of foreign language vocabulary acquisition,' *Language Learning* 43/4: 559–617.

Ellis, N. and **L. Collins.** 2009. 'Input and second language acquisition: the roles of frequency, form, and function introduction to the special issue,' *Modern Language Journal* 93: 329–35.

Ellis, R. 1995. 'Modified oral input and the acquisition of word meanings,' *Applied Linguistics* 16: 409–35.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Field, A. 2005. *Discovering Statistics Using SPSS*. Sage Publications.

Gee, N. R., D. L. Nelson, and **D. Krawczyk.** 1999. 'Is the concreteness effect a result of underlying network interconnectivity?,' *Journal of Memory and Language* 40: 479–97.

Gilhooly, K. J. and **R. H. Logie.** 1980. 'Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words,' *Behavior Research Methods & Instrumentation* 12/4: 395–427.

Godfrey, J. J. and **E. Holliman.** 1993. *Switchboard-1* [CD-ROM]. Linguistic Data Consortium.

Graesser, A. C., D. S. McNamara, M. M. Louwerse, and **Z. Cai.** 2004. 'Coh-Metrix: analysis of text on cohesion and language,' *Behavioral Research Methods, Instruments, and Computers* 36: 193–202.

Henriksen, B. 1999. 'Three dimensions of vocabulary development,' *Studies in Second Language Acquisition* 21: 303–17.

Higgins, D., X. Xi, K. Zechner, and **D. Williamson.** 2011. 'A three-stage approach to the automated scoring of spontaneous spoken responses,' *Computer Speech and Language* 25/2: 282–306.

Hoey, M. 2005. *Lexical Priming: A New Theory of words and Language*. Routledge.

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

Kline, P. 1999. *The Handbook of Psychological Testing*. Routledge.

Landauer T. K., D. S. McNamara, S. Dennis, and **W. Kintsch** (eds). 2007. *LSA: A Road to Meaning*. Erlbaum.

Laufer, B. and **P. Nation.** 1995. 'Vocabulary size and use: Lexical richness in L2 written production,' *Applied Linguistics* 16/3: 307–22.

Lennon, P. 1996. 'Getting 'easy' verbs wrong at the advanced level,' *International Review of Applied Linguistics in Language Teaching* 34/1: 23–36.

Levenston, E. and **S. Blum.** 1977. 'Aspects of lexical simplification in the speech and writing of advanced adult learners' in P. S. Corder and E. Roulet (eds): *The Notions of Simplification, Interlanguages and Pidgins and their Relation to Second Language Pedagogy*. Librairie Droz, pp. 51–72.

Malvern, D., B. J. Richards, N. Chipere, and **P. Duran.** 2004. *Lexical Diversity and Language Development. Quantification and Assessment*. Palgrave Macmillan.

Manchon, R. M., L. Murphy, and **J. Roca de Larios.** 2007. 'Lexical retrieval processes and strategies in second language writing: A synthesis of empirical research,' *International Journal of English Studies* 7: 147–72.

McCarthy, P. M. and **S. Jarvis.** 2010. 'MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment,' *Behavior Research Methods* 42: 381–92.

Meara, P. 1996. 'The dimensions of lexical competence' in G. Brown, K. Malmkjaer, and J. Williams (eds): *Performance and Competence in Second Language Acquisition*. Cambridge University Press, pp. 35–53.

Meara, P. 2005a. 'Designing vocabulary tests for English, Spanish and other languages' in C. Butler, S. Christopher, M. Á. Gómez González, and S. M. Doval-Suárez (eds): *The Dynamics of Language Use*. John Benjamins Press, pp. 271–85.

Meara, P. M. 2005b. 'Lexical frequency profiles: A Monte Carlo analysis,' *Applied Linguistics* 26/1: 32–47.

Moon, R. 1992. 'Textual aspects of fixed expressions in learners' dictionaries' in J. A. Pierre and B. Henri (eds): *Vocabulary and Applied Linguistics*. Macmillan, pp. 13–27.

Multon, K. 2010. 'Interrater reliability' in N. Salkind (ed.): *Encyclopedia of Research Design*. Sage Publications, Inc, pp. 627–9.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, I. S. P. 2007. 'Fundamental issues in modelling and assessing vocabulary knowledge' in H. Daller, J. Milton, and J. Treffers-Daller (eds): *Modelling and Assessing Vocabulary Knowledge*. Cambridge University Press, pp. 33–43.

Nesselhauf, N. 2003. 'The use of collocations by advanced learners of English and some implications for teaching,' *Applied Linguistics* 24: 223–42.

Newman, M. L., C. J. Groom, L. D. Handelman, and **J. W. Pennebaker.** 2008. 'Gender difference in language use: An analysis of 14,000 text samples,' *Discourse Processes* 45: 211–36.

Paivio, A. 1991. 'Dual coding theory: Retrospect and current status,' *Canadian Journal of Psychology* 45: 255–287.

Paivio, A., J. C. Yuille, and **S. Madigan.** 1968. 'Concreteness, imagery, and meaningfulness values for 925 nouns,' *Journal of Experimental Psychology Monograph Supplement* 76: 1–25.

Qian, D. D. 1999. 'Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension,' *Canadian Modern Language Review* 56: 282–308.

Read, J. 1998. 'Validating a test to measure depth of vocabulary knowledge' in A. Kunnan (ed.): *Validation in Language Assessment*. Lawrence Erlbaum, pp. 41–60.

Read, J. 2000. *Assessing Vocabulary*. Cambridge University Press.

Rosch, E., C. Mervis, W. Gray, D. Johnson, and **P. Boyes-Braem.** 1976. 'Basic objects in natural categories,' *Cognitive Psychology* 8: 573–605.

Salsbury, T., S. A. Crossley, and **D. S. McNamara.** 2011. 'Psycholinguistic word information in second language oral discourse,' *Second Language Research* 27: 343–60.

Schmitt, N. 1998. 'Tracking the incremental acquisition of a second language vocabulary: A longitudinal study,' *Language Learning* 48/2: 281–317.

Skehan, P. 1998. *A Cognitive Approach to Language Learning*. Oxford University Press.

Tabachnick, B. G. and **L. S. Fidell.** 2001. *Using Multivariate Statistics, 4th edn*. Allyn & Bacon.

Toglia, M. P. and **W. F. Battig.** 1978. *Handbook of Semantic Word Norms*. Lawrence Erlbaum Associates.

Whitten, I. A. and **E. Frank.** 2005. *Data Mining*. Elsevier.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Wray, A. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford University Press.

Zareva, A., P. Schwanenflugel, and **Y. Nikolova.** 2005. 'Relationship between lexical competence and language proficiency: Variable sensitivity,' *Studies in Second Language Acquisition* 27: 567–95.

Zipf, G. K. 1935. *The Psycho-biology of Language*. Houghton-Mifflin.

NOTES ON CONTRIBUTORS

Scott Crossley is an Associate Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, cognitive science, discourse processing, and discourse analysis. His primary research focuses on corpus linguistics and the application of computational tools in second language learning and text comprehensibility. *Address for correspondence*: Scott Crossley, Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA, USA. <*sacrossley@gmail.com*>

Tom Salsbury is an Associate Professor in the Department of Teaching and Learning, College of Education, at Washington State University. He has published in the areas of pragmatics and language learning, discourse analysis, and content-based instruction. His current work is in vocabulary development and the analysis of longitudinal, oral, and written texts produced by second language learners. *Address for correspondence*: College of Education, Washington State University, Pullman, WA, USA. <tsalsbury@wsu.edu>

Danielle McNamara is a Professor at Arizona State University and Senior Research Scientist at the Learning Sciences Institute. 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. *Address for correspondence*: Learning Sciences Institute, Arizona State University, Tempe, AZ, USA. <Danielle.Mcnamara@asu.edu>