tesol QUARTERLY

Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application

KRISTOPHER KYLE AND SCOTT A. CROSSLEY
Georgia State University
Atlanta, Georgia, United States

This study explores the construct of lexical sophistication and its applications for measuring second language lexical and speaking proficiency. In doing so, the study introduces the Tool for the Automatic Analysis of LExical Sophistication (TAALES), which calculates text scores for 135 classic and newly developed lexical indices related to word frequency, range, bigram and trigram frequency, academic language, and psycholinguistic word information. TAALES is freely available; runs on Windows, Mac, and Linux operating systems; and has a simple graphic user interface that allows for batch processing of .txt files. The tool is fast, reliable, and outputs results to a comma-separated value file that can be accessed using spreadsheet software. The study examines the ability of TAALES indices to explain the variance in human judgments of lexical proficiency and speaking proficiency for second language (L2) learners. Overall, these indices were able to explain 47.5% of the variance in holistic scores of lexical proficiency and 48.7% of the variance in holistic scores of speaking proficiency. This study has important implications for second language acquisition, for assessing L2 learners' productive skills (writing and speaking), and for L2 pedagogy. Limitations and future directions are also discussed.

doi: 10.1002/tesq.194

In the past decade, the use of computer-aided automatic text analysis has grown in use, largely due to the introduction of a number of computational tools that are either freely available or relatively inexpensive. Tools such as VocabProfile (Cobb, 2013; Heatley, Nation, & Coxhead, 2002), Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007), and Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014) have been made available to second language (L2) educators and researchers. Such tools have proven valuable in studying L2 vocabulary acquisition (e.g., Crossley, Salsbury, & McNamara, 2010), reading difficulty (e.g., Cobb, 2007; Crossley, Greenfield, & McNamara, 2008), writing quality (Crossley & McNamara, 2012; McNamara, Crossley, & McCarthy, 2010), speaking proficiency (e.g., Crossley, Clevinger, & Kim, 2014; Crossley & McNamara, 2013) and lexical proficiency (e.g., Crossley, Salsbury, McNamara, & Jarvis, 2011a). While these tools have clearly broadened our understanding of a number of different linguistic and pedagogical aspects of L2 learning, they are limited in the number and variety of lexical inquiries available.

For example, Coh-Metrix provides only frequency information from the CELEX database, and VocabProfile provides only frequency information from the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). Although these frequency measures and the other available lexical measures have made significant contributions to lexical studies, many new and potentially important research measures such as SUBTLEXus frequency counts (Brysbaert & New, 2009), bigram and trigram frequency indices (Crossley, Cai, & McNamara, 2012), the Academic Formulas List (AFL; Simpson-Vlach & Ellis, 2010), and updated psycholinguistic word and bigram norms such as those collected by Brysbaert, Warriner, and Kuperman (2013) and Kuperman, Stadthagen-Gonzales, and Brysbaert (2012) are not present in current tools.

In addition to limits with the indices calculated by the currently available tools, functionality is also a confining factor. The freely available online version of Coh-Metrix, for example, despite reporting a number of lexical indices, can process only one text at a time, practically limiting the number of texts that can be examined. VocabProfile, despite being very easy to use, also allows files to be processed only one at a time. LIWC, on the other hand, is able to very quickly process texts in batches and allows the addition of word categories. Despite these advantages, however, LIWC is unable to deal with reference corpus frequency, range counts, or psycholinguistic norms such as concreteness or imageability. It is also not freely available.

The current study seeks to help address the lexical limitations of these tools by introducing the Tool for the Automatic Analysis of LExical Sophistication (TAALES). TAALES is freely available and easy to use, works on most operating systems (Windows, Mac, Linux), allows for batch processing of text files, and incorporates more than 130 classic and recently developed indices of lexical sophistication. In this article, we introduce TAALES and examine its validation through a number of analyses that investigate how indices of lexical sophistication can predict written lexical proficiency and speaking proficiency. We discuss how these analyses and, by proxy, TAALES can be used to inform L2 assessment and pedagogy.

Lexical Sophistication

Although an exact definition of lexical sophistication has yet to be agreed upon, the construct of lexical sophistication involves both the depth and breadth of lexical knowledge available to speakers, readers, and writers (Meara, 1996, 2005a; Read, 1998). A number of indices to measure the depth and breadth of lexical sophistication in L2 learners have been proposed. Perhaps the most well-attested method of measuring lexical sophistication is to use a corpus-derived frequency count (e.g., Laufer & Nation, 1995; Crossley et al., 2010; Crossley and McNamara, 2013) to examine the frequency of words in a text. There are a variety of methods to assess frequency, with the most common method likely being the examination of frequency bands (e.g., Laufer, 1994; Laufer & Nation, 1995; Morris & Cobb, 2004). Another common approach is to measure frequency based on word counts taken from representative corpora (e.g., McNamara et al., 2010). Less common, but potentially as important as other measures of frequency, are indices based on frequency ranges, which provide a count of in how many documents in a corpus a lexical item occurs (e.g., Gries, 2008; Crossley, Subtirelu, & Salsbury, 2013). In addition to word frequency counts, researchers have also begun to examine how bigram and trigram frequencies are linked to lexical sophistication (e.g., Crossley et al., 2012). Academic lists, such as the Academic Word List (AWL; Coxhead, 2000) and the Academic Formulas List (Simpson-Vlach & Ellis, 2010), have also been noted in the literature as important indicators of lexical sophistication for learners in academic contexts (see Coxhead, 2011, for an overview of the impact of the AWL). Finally, psycholinguistic properties of words, which have been of interest in the field of cognitive science for some time (e.g., Coltheart, 1981), have also been tied to lexical sophistication (e.g., Crossley et al., 2010, 2011a; Crossley, Salsbury, McNamara, & Jarvis, 2011b; McNamara et al., 2010). We discuss each of these types of lexical sophistication below.

Frequency

Frequency as a measure of lexical sophistication is based on the notion that words that are more frequent in natural language data are learned earlier and used more often than words that are less frequent in natural language data, although this may not be true in all cases (Crossley et al., 2010, 2014). Frequency has been shown to affect lexical decision times (Kuperman et al., 2012), indicating that high-frequency words are processed more quickly than low-frequency words. In the areas of first and second language writing and second language speaking, word frequency indices are predictive of holistic quality/proficiency scores (e.g., Crossley, Cobb, & McNamara, 2013; Laufer & Nation, 1995) with essays and speaking samples that include higher frequency words tending to earn lower quality and/or proficiency scores.

In writing and speaking assessment studies, two main types of frequency measures have been used. The first is based on frequency bands (e.g., Laufer, 1994; Laufer & Nation, 1995; Morris & Cobb, 2004) wherein word frequency scores are derived by rank-ordering words in a frequency list (e.g., the most frequent word in a corpus would receive a score of 1), and then categorized based on whether they are in the most frequent 1,000 words (1K list), 2,000 words (2K list), and so on. The percentage of the text that occurs in a particular band can then be used as a predictor variable. Such counts can be obtained by the online tool VocabProfile (Cobb, 2013; Heatley et al., 2002). The other main approach is to determine the reference corpus (e.g., BNC Consortium, 2007) frequency of each word that occurs in a target text and create an average frequency score for a text by dividing the sum of all frequency scores by the number of words in a text.

Although both methods of determining frequency information discussed above have demonstrated the ability to predict holistic scores of quality and proficiency in L2 learners, the latter seems to have a number of advantages. First, frequency counts such as those provided by Coh-Metrix provide a finer grained account of frequency and thus may be better suited for measuring both large and small gains in an individual's productive vocabulary (see, e.g., Meara, 2005b). In addition, when used side by side, frequency count measures have been shown to produce predictor models that are more accurate than predictor models that include only frequency band measures (Crossley, Cobb, et al., 2013).

Range

Because frequency measures are based on counts of how many times a word or a word family occurs in a corpus as a whole, they are not sensitive to how widely a particular word or word family is used. Range (also referred to as *dispersion, entropy,* and *contextual diversity*) measures account for how widely a word or word family is used, usually by providing a count of the number of documents in which that word occurs. To illustrate, in the written portion of the BNC, the words *cent, next,* and *four* have very similar frequencies of occurrence (34,367, 34,590, and 34,290 occurrences, respectively), preliminarily indicating similar importance. Their range values, however, are quite different: *Cent* only occurs in approximately half of the documents in the BNC (1,526 documents, or 49.50% of documents), whereas *next* and *four* occur in almost 90% of the documents in the BNC (2,753 and 2,754 documents, or 89.30% and 89.33% of documents, respectively), suggesting that *next* and *four* may be more common across texts and thus less sophisticated than *cent*. Gries (2008) has stressed the importance of range in corpus linguistics, and range indices have been used to successfully distinguish frequent verbs produced by L2 speakers of English from frequent verbs produced by native English speakers (Crossley, Subtirelu, et al., 2013).

N-gram Frequencies

Many researchers have begun to shift their focus from single lexical items to larger units (e.g., Biber, Conrad, & Cortes, 2004). Recently, this focus has extended to natural language processing (NLP) applications, and n-grams (combinations of *n* number of words, such as the bigram *other hand*) have proven useful in areas such as native language identification/crosslinguistic influence (e.g., Jarvis & Crossley, 2012; Kyle, Crossley, Dai, & McNamara, 2013) and writing quality (Crossley et al., 2012). Crossley et al. (2012), for example, calculated a variety of frequency-based indices for bigrams and trigrams in written and spoken subsections in the BNC. Indices that measured bigram frequency, bigram proportions, and bigram accuracy were predictive of human judgments of essay quality.

Academic Lists

Academic word lists comprise words that occur relatively infrequently in general language corpora, but occur frequently in academic texts. Xue and Nation (1984), for example, developed the University Word List (UWL), an academic word list comprising 836 word families that, when added to the General Service List (GSL), provided 85.9% coverage of academic texts. Later, Coxhead (2000) created the Academic Word List, which had improved coverage and used fewer word families than the UWL. The AWL is based on an academic corpus comprising journal articles and textbook chapters from a total of 28 subject areas in four broad disciplines. Since 2000, the AWL has been very influential in English for academic purposes (EAP) research (Coxhead, 2011).

EAP researchers have also investigated multiword units that are pervasive in academic language. Simpson-Vlach and Ellis (2010), for example, developed the Academic Formulas List (AFL). The AFL is based on spoken and written academic corpora. The spoken corpus comprised the Michigan Corpus of Academic Spoken English (MICASE, 1.9 million words; Simpson, Briggs, Ovens, & Swales, 2002) and academic speaking samples from the BNC (431,000 words). The written corpus comprised Hyland's (2004) corpus of journal articles (1.2 million words) and a 931,000-word subset of academic writing from the BNC. Formulas were chosen if they occurred more frequently in the academic corpora than in nonacademic spoken and written corpora. Additionally, for formulas to be included in the final lists, they had to occur in at least four of five spoken genres, three of four written genres, or six of nine combined genres. These cutoff points resulted in three large lists of three-, four-, and five-word formulas: AFL spoken, AFL written, and AFL core.

Psycholinguistic Properties of Words

The psycholinguistic properties of words have been of interest to researchers in both cognitive science (Coltheart, 1981; Toglia & Battig, 1978) and second language acquisition (e.g., Crossley, Weston, McLain Sullivan, & McNamara, 2011). These properties have been used to explain the variance in lexical decision times (e.g., Kuperman et al., 2012), holistic scores of writing quality (e.g., Crossley & McNamara, 2012), lexical proficiency (e.g., Crossley et al., 2011a) and speaking proficiency (e.g., Crossley & McNamara, 2013). Of particular interest have been the psycholinguistic properties of concreteness, familiarity, imageability, meaningfulness, and age of acquisition, each of which is outlined below.

Concreteness. Concreteness ratings are based on perceptions of how abstract a word is based on how simple it is to describe the meaning of that word. If one can describe a word by simply pointing to the object it signifies, such as the word *apple*, a word can be said to be concrete, while if a word can be explained only using other words, such as *infinity* or *impossible*, it can be considered more abstract (Brysbaert et al., 2013).

Familiarity. Familiarity scores are based on judgments of how familiar words are to adults and are correlated with frequency counts (Crossley, Subtirelu, et al., 2013). Words such as *breakfast*, *radio*, and *book* all have high familiarity scores, whereas words such as *encephalon* and *egress* have low familiarity scores.

Imageability. Imageability scores are based on judgments of how easy it is to create an image of a word. A word such as *beach* is highly imageable, whereas a word such as *philology* is not very imageable.

Meaningfulness. Meaningfulness scores are based on judgments of how related a word is to other words. The word *beautiful*, for example, has a very high meaningfulness score, whereas the word *adze* has a low meaningfulness score.

Age of acquisition. Age of acquisition (AoA) indices are based on human judgments of the age at which a particular word is learned. AoA values are positively correlated with both response times and accuracy scores on lexical decision tasks, and they add to the variance in these that are explained by traditional measures (e.g., frequency; Kuperman et al., 2012).

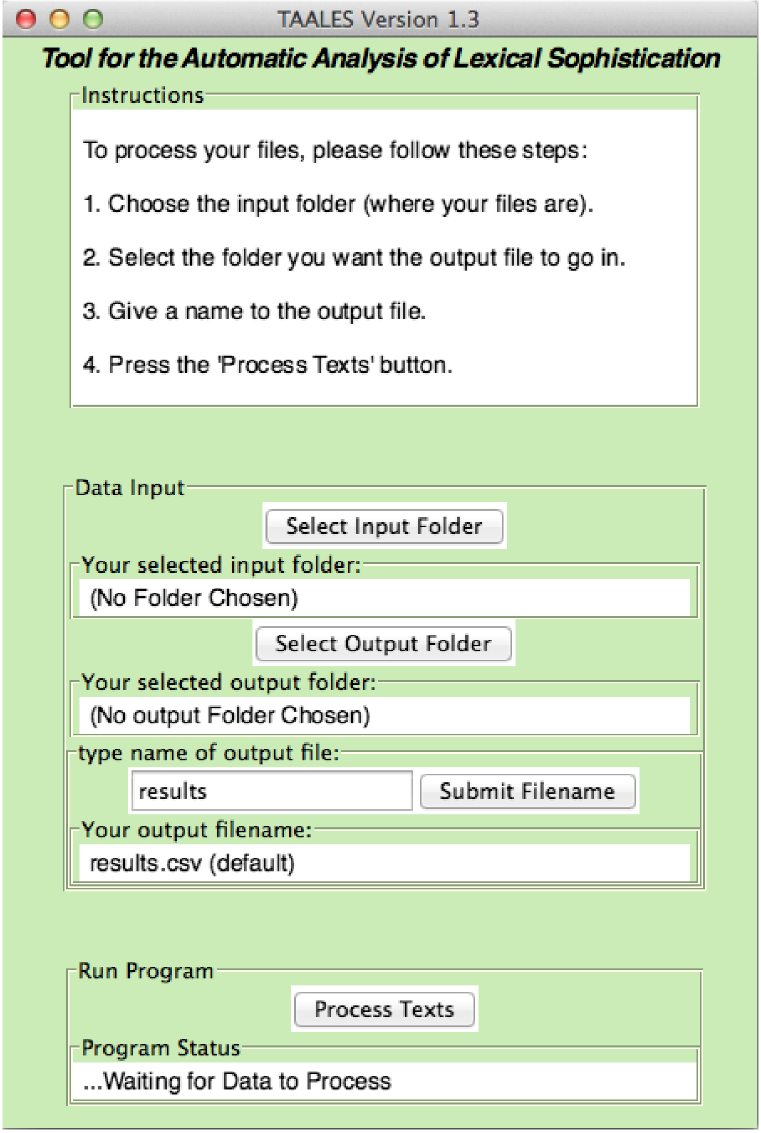
Accessibility of Lexical Sophistication Indices

Although indices of lexical sophistication have been demonstrated to be important indicators of L2 language proficiency, reading proficiency, lexical proficiency, speaking proficiency, and writing proficiency (e.g., Crossley & McNamara, 2013; Crossley et al., 2011a), all of which have important applications in the classroom (Laufer, 1994), calculating many of the indices outlined above is beyond the reach of many researchers and language teaching professionals who may lack programming knowledge. Where user-friendly tools exist, such as VocabProfile (Cobb, 2013; Heatley et al., 2002) and Coh-Metrix (Graesser et al, 2004; McNamara & Graesser, 2012; McNamara et al., 2014), they may be impractical for processing large numbers of texts and may be limited in the number and variety of indices they report on. In the current article, we seek to address this gap by introducing and validating TAALES, a freely available, cross-platform, user-friendly tool for analyzing more than 130 indices of lexical sophistication. The program is accessed through a graphical user interface (see Figure 1), processes .txt formatted files, and produces output in a comma-separated value (.csv) file.

TAALES includes a number of indices that focus on the five areas of lexical sophistication discussed above: lexical frequency, range, n-gram frequency, academic vocabulary, and psycholinguistic word properties.[1] In addition to introducing TAALES, this article investigates the validity of the included indices by examining how the indices predict variance in holistic judgments of lexical proficiency and speaking proficiency. The validation of TAALES is guided by the following research questions:

[1] Recent studies (e.g., McCarthy & Jarvis, 2010) indicate that indices of lexical diversity are likely related to text cohesion (i.e., word repetition). For this reason, TAALES does not include lexical diversity indices.

1. Can indices of frequency, range, academic vocabulary use, and lexical psycholinguistic properties reported by TAALES account for the variance in holistic judgments of lexical proficiency?
2. Can indices of frequency, range, academic vocabulary use, and lexical psycholinguistic properties reported by TAALES account for the variance in holistic judgments of speaking proficiency?


TAALES Version 1.3

Tool for the Automatic Analysis of Lexical Sophistication

Instructions

To process your files, please follow these steps:

1. Choose the input folder (where your files are).
2. Select the folder you want the output file to go in.
3. Give a name to the output file.
4. Press the 'Process Texts' button.

Data Input

Your selected input folder: (No Folder Chosen)

Your selected output folder: (No output Folder Chosen)

type name of output file: results

Your output filename: results.csv (default)

Run Program

Program Status: ...Waiting for Data to Process

FIGURE 1. Screenshot of the TAALES user interface

GENERAL METHODS

To assess the validity of the indices of lexical sophistication reported by TAALES, we investigated the relationship between the indices (outlined in greater detail below) and two types of language proficiency scores.

Overview of Corpora

Our two corpora were a corpus of unstructured free-writes written by English language learners and native English speakers that had been scored for holistic lexical proficiency (Crossley et al., 2011a) and a corpus of independent TOEFL speaking samples that had been given holistic speaking proficiency scores (Crossley & McNamara, 2013). In each case, trained human raters assigned a holistic score between 2 and 8 (in the case of speaking proficiency) and between 1 and 5 (in the case of lexical proficiency). In each case, interrater reliability between raters was acceptable. Descriptive statistics for the corpora are presented in Table 1. These are discussed in greater detail below.

Lexical proficiency corpus. The holistic lexical proficiency scores used in this study were collected for and reported on in Crossley et al. (2011a). Crossley et al. collected 180 unstructured writing samples from 10 L2 English learners at an intensive program over a 1-year period. The number of texts collected from each learner was not equal because not all learners submitted the same number of writing samples. The texts were grouped by the participant's institutional TOEFL scores into beginning ($n = 60$), medium ($n = 60$), and high ($n = 60$) proficiency samples. These samples were augmented with 60 unstructured writing samples from undergraduate native speakers leading to a total corpus of $N = 240$. The writing samples were holistically scored for lexical proficiency using a carefully prepared rubric by expert raters, and interrater reliability was acceptable ($r = .796$).

Speaking proficiency corpus. Holistic speaking proficiency scores are derived from the TOEFL public use data set, which includes

TABLE 1
Descriptive Statistics for the Two Corpora With Holistic Scores

Corpus	*N*	Min. score	Max. score	Mean score	*SD* score
Lexical proficiency	240	1.333	5.000	3.215	1.050
Speaking proficiency	244	2.000	8.000	5.266	1.457

data from actual administrations of the TOEFL during 2007. Unlike the lexical proficiency corpus, the speaking proficiency corpus contained no native speaker samples. The speaking section of the TOEFL includes six speaking tasks, two of which are *independent* and four of which are *integrated*. The independent tasks, which are used in this study, ask test takers to give an opinion on a particular topic or choose one side of controversial topic and defend that choice. Test takers are given 15 seconds to prepare their response for each task and then have 45 seconds to respond. Responses are scored based on a 4-point rubric by ETS-trained raters.

Responses from 244 participants were orthographically transcribed and saved in .txt files. Due to the relatively short length of the responses (many are under 100 words), the two independent responses from each test taker were combined, and scores from each of the responses were added together (see Crossley & McNamara, 2013).

Indices of Lexical Sophistication

For all indices of lexical sophistication, only words/bigrams/trigrams in the target text that are included in the appropriate database are used to calculate index scores. For normed indices, scores are calculated by dividing the sum of the scores by the number of words/bigrams/trigrams in the target text that have entries in the appropriate database.

Word frequency indices. Unless otherwise noted, frequency scores are calculated in TAALES by dividing the sum of the frequency scores for the tokens in a text by the number of tokens in that text that received a frequency score (i.e., words in the text that are not in the selected frequency lists are not included in the frequency counts). For each frequency list included in TAALES, we calculate scores for all words (AW), content words (CW), and function words (FW). In addition, we provide logarithmic transformations for each of these indices to control for Zipfian effects common in word frequency lists. Frequency indices were calculated from the following frequency lists:

Thorndike-Lorge. The Thorndike-Lorge (TL) written frequency counts are based on Lorge's 4.5 million-word corpus of popular magazine articles (Thorndike & Lorge, 1944). Frequency counts are included for the 14,865 lemmas found in the MRC psycholinguistic database (Coltheart, 1981).

Kucera-Francis written frequency. Kucera-Francis (KF) written frequencies are based on the Brown corpus (Kucera & Francis, 1967), which consists of approximately 1 million words of texts published in the United States in 1961. Frequency counts are available for the 17,862 lemmas found in the MRC psycholinguistic database (Coltheart, 1981).

Brown verbal frequency. Brown verbal frequencies (Brown, 1984) are based on the 1-million word London-Lund Corpus of English Conversation (Svartvik & Quirk, 1980). TAALES includes frequency counts for the 8,126 lemmas found in the MRC psycholinguistic database (Coltheart, 1981).

British National Corpus. The British National Corpus, version 3 (BNC XML ed.) (BNC Consortium, 2007) comprises 100 million words of written (90 million words) and spoken (10 million words) English from Great Britain. TAALES includes frequency scores for an 80-million word subset of the written section and the 10-million word spoken section of the BNC. TAALES provides separate scores for written frequency (128,793 entries) and spoken frequency (25,953 entries). All frequency counts from the BNC are raw (not lemmatized).

SUBTLEXus. The SUBTLEXus corpus (Brysbaert & New, 2009) comprises subtitles from 8,388 films and television series from the United States. The corpus includes 51 million words. TAALES includes un-lemmatized frequency norms for 74,286 words.

Range indices. In addition to frequency information, TAALES includes a number of range indices. Range indices are calculated for the spoken (574 texts) and written (3,083 texts) subsets of the BNC, SUBTLEXus (8,388 texts), and Kucera-Francis (500 texts) frequency lists. Also included is a range count based on Kucera and Francis's (1967) 15 text categories, which can be described, roughly, as genres.

N-gram indices. Crossley et al. (2012) calculated bigram and trigram frequencies from both written (80 million words) and spoken subcorpora (10 million words) of the BNC. TAALES utilizes these frequencies to calculate five types of indices. The first type of index includes non-normalized logarithm-transformed frequency counts. The second involves dividing the sum of the bigram/trigram frequency scores in a text with the number of bigrams/trigrams present in the text that are represented in the reference corpus (i.e., n-gram frequency by n-gram). The third type of index involves dividing the sum of the bigram/trigram frequency scores with the number of words in the target text (i.e., n-gram frequency by word). In addition to the frequency scores, bigram and trigram type counts were calculated (i.e., the number of unique bigrams and trigrams per text) as well as proportion scores. The proportion scores were calculated by dividing the number of unique bigrams/trigrams in the text that are represented in the reference corpus by the number of words in the text. For more information, see Crossley et al. (2012).

Academic list indices. In addition to traditional frequency counts and n-gram counts, TAALES includes word and n-gram level academic lists. These indices are calculated by counting the number of tokens in the text that occur in an academic list (raw scores) and dividing by the number of words in the text (normed scores). These indices are calculated from the AWL and the AFL.

Word information indices. Word information scores are derived from the MRC Psycholinguistic Database (Coltheart, 1981), Brysbaert et al. (2013), and Kuperman et al. (2012). Word information scores are calculated by adding the score for each token in a text that is given a word information score and dividing by the number of tokens in the text that is given a word information score (i.e., words that are not in the selected word information lists but are in the texts are not included in the counts). Word information indices are calculated for all words, content words, and function words. Word information indices were calculated from the following lists:

Familiarity. TAALES includes familiarity scores for 4,943 lemmas from the MRC psycholinguistics database (Coltheart, 1981).

Concreteness. TAALES includes concreteness ratings for 4,315 lemmas from the MRC psycholinguistics database (Coltheart, 1981). Also included are concreteness ratings for 37,058 lemmas and 2,896 bigrams collected by Brysbaert et al. (2013) via Amazon Mechanical Turk.

Imageability. TAALES includes imageability ratings for 4,848 lemmas from the MRC psycholinguistic database (Coltheart, 1981).

Meaningfulness. TAALES includes an index based on the meaningfulness norms for 2644 lemmas collected by Toglia and Battig

(1978) and included in the MRC psycholinguistic database (Coltheart, 1981).

Age of acquisition. TAALES includes an AoA index based on the Kuperman et al. (2012) AoA list, which includes 30,121 lemmas that were judged for AoA using Amazon Mechanical Turk.

Analysis

For each of the two corpora (lexical proficiency and speaking proficiency) we first determined whether the proposed indices of lexical sophistication were distributed normally. Any indices that were not roughly normal were discarded. Correlations were then run to determine whether there was a statistical ($p < .05$) and meaningful (at least a small effect size, $r > .1$) relationship between the proposed indices and the holistic scores for a particular corpus. Multicollinearity was then checked. Any indices that were highly collinear ($r \geq .9$) were flagged; the index with the strongest correlation with holistic scores was retained, and the other indices were removed. Additionally, any indices that were very strongly correlated ($r \geq .9$) with word count were removed to control for text length effects, which strongly influence human judgments of linguistic proficiency (Ferris, 1994). The remaining indices were included as predictor variables in a stepwise multiple regression to determine whether the proposed indices of lexical sophistication could account for the variance in each type of holistic score. Multiple regression analysis has been used in a great number of studies to model holistic scores (e.g., Deane & Quinlan, 2010). In order to ensure that the resulting multiple regression model was stable across the entire data set and generalizable to populations outside the selected data set, a 10-fold *leave one out cross validation* (see Witten, Frank, & Hall, 2011, for a detailed description of this procedure) multiple regression was conducted using the predictor variables chosen by the initial stepwise multiple regression.

RESULTS

Study 1: Lexical Proficiency

Of the 135 indices considered, 42 violated the assumption of normality and were removed from further consideration. Of the remaining 93 indices, 70 were statistically correlated ($p < .05$) with the holistic scores of lexical proficiency. One of these, Bigram Type Count, was highly correlated ($r > .9$) with word count, and was removed. After checking for multicollinearity, 40 indices remained (see the Appendix) and were entered into a stepwise multiple regression to determine whether the proposed indices of lexical sophistication could predict the variance in holistic lexical proficiency scores. The regression produced a model that included five indices (BNC Spoken Bigram Proportion, BNC Written Range AW, BNC Written Trigram Frequency Normed [word] Logarithm, MRC Familiarity CW, and MRC Meaningfulness CW) and explained 51.7% ($r = .719$, $r^2 = .517$) of the variance in holistic lexical proficiency scores. See Table 2 for a summary of the stepwise multiple regression models and for the beta weights for each index in the final model.

To validate the model, a 10-fold cross validation (CV) multiple regression was also conducted using the five variables identified in the previous multiple regression model. The 10-fold CV regression models explained 47.5% of the variance in holistic lexical proficiency scores ($r = .689$, $r^2 = .475$), indicating that the five-variable model identified in the initial stepwise multiple regression is stable across subsections of the data set and generalizable to other populations.

Study 2: Speaking Proficiency

Of the 135 indices considered, 34 violated the assumption of normality and were removed from further consideration. Of the remaining 101 indices, 44 were statistically correlated ($p < .05$) with the holistic scores of speaking proficiency. Four of these (Bigram Type Count, Trigram Type Count, BNC Bigram Written Frequency Logarithm, and BNC Bigram Spoken Frequency Logarithm) were highly correlated ($r > .9$) with word count and were removed. After checking for multicollinearity, 30 indices remained (see the Appendix) and

TABLE 2
Summary of Stepwise Multiple Regression Models for Lexical Proficiency

Entry	Predictors included	r	R^2	R^2 change	β	SE	B
1	BNC Spoken Bigram Proportion	.410	.168	.168	5.781	.699	.528
2	BNC Written Range AW	.653	.426	.259	−0.159	.018	−.557
3	BNC Written Trigram Frequency Normed (word) Logarithm	.694	.482	.055	2.846	.658	.272
4	MRC Familiarity CW	.707	.501	.019	−0.021	.008	−.145
5	MRC Meaningfulness CW	.719	.517	.016	−0.007	.003	−.132

Note. Estimated constant term = 28.196; β = unstandardized beta; *SE* = standard error; *B* = standardized beta.

were entered into a stepwise multiple regression to determine whether the proposed indices of lexical sophistication could predict the variance in holistic scores of speaking proficiency. If beta weights of any indices in the multiple regression model showed signs of suppression (i.e., a noise error in a predictor variable is suppressed by a second predictor variable increasing the *r* value in the model; Cohen & Cohen, 1983) as evidenced by a switched sign, the model was rerun without that variable. For this analysis, the first three models had sign-switched variables and had to be rerun (a total of five variables were removed). The final model, including five indices (BNC Trigram Written Frequency Logarithm, SUBTLEXus Range CW Logarithm, ALL AWL RAW, MRC Familiarity AW, and Brown Frequency CW Logarithm) was able to explain 51.9% (r = .720, r^2 = .519) of the variance in holistic speaking proficiency scores. See Table 3 for a summary of the stepwise multiple regression models and for the beta weights for each index in the final model.

To validate the model, a 10-fold CV multiple regression was also conducted using the five variables identified in the previous multiple regression model. On average, the 10-fold CV regression models explained 48.7% of the variance in holistic speaking proficiency scores (r = .698, r^2 = .487), indicating that the five-variable model identified in the initial stepwise multiple regression is stable across subsections of the data set and generalizable to other populations.

DISCUSSION

Study 1: Lexical Proficiency

Of the 135 original indices, 40 unique (e.g., not strongly correlated with any other indices) indices were correlated with holistic scores of

TABLE 3
Summary of Stepwise Multiple Regression Models for Speaking Proficiency

Entry	Predictors included	*r*	R^2	R^2 change	β	*SE*	*B*
1	BNC Trigram Written Frequency Logarithm	.593	.352	.352	0.031	.003	.574
2	SUBTLEXus Range CW Logarithm	.669	.448	.096	−3.587	.892	−.239
3	All AWL Raw	.690	.476	.028	0.104	.025	.218
4	MRC Familiarity AW	.708	.501	.025	−0.0805	.025	−.164
5	Brown Frequency CW Logarithm	.720	.519	.018	1.790	.608	.172

Note. Estimated constant term = 60.453; β = unstandardized beta; *SE* = standard error; *B* = standardized beta.

lexical proficiency at a statistical ($p < .05$) and meaningful (r > .1) level. Five of these indices (BNC Spoken Bigram Proportion, BNC Written Range AW, BNC Written Trigram Frequency Normed [word] Logarithm, MRC Familiarity CW, and MRC Meaningfulness CW) were able to explain 51.7% of the variance in holistic lexical proficiency scores for the training set, validating their use for measuring lexical sophistication. The index that accounted for the largest percentage of the variance (25.9%) in holistic scores of lexical sophistication was the BNC Written Range index for all words. These range scores were negatively correlated with lexical proficiency scores, indicating that words that are used in fewer contexts are judged to be more sophisticated than those that are widely used. The percentage of bigrams in the target text that also were relatively frequent in the BNC spoken corpus index also accounted for a relatively large percentage of the variance (16.8%) in lexical sophistication scores. Writing samples that demonstrate lexical proficiency tend to have a higher percentage of bigrams that also occur in the spoken portion of the BNC, indicating that collocation knowledge is an important aspect of lexical proficiency. Furthermore, the normed frequency of trigrams (based on the written portion of the BNC) was positively correlated with holistic scores of lexical proficiency and explained 5.5% of the variance in the regression model, indicating that writing samples containing more frequent trigrams are judged to be more lexically sophisticated than those with lower frequency trigrams. Finally, the MRC familiarity and meaningfulness scores for content words were negatively correlated with holistic scores of lexical proficiency, indicating that words that are less familiar and less meaningful are judged to be more lexically sophisticated. These indices explained a small amount of the variance in the regression model (accounting for 1.9% and 1.6% of the variance, respectively).

Study 2: Speaking Proficiency

Of the 135 indices initially considered, 30 unique indices were statistically ($p < .05$) and meaningfully ($r > .1$) correlated with holistic scores of speaking proficiency. Five of these indices (BNC Trigram Written Frequency Logarithm, SUBTLEXus Range CW Logarithm, All AWL Raw, MRC Familiarity AW, and Brown Frequency CW Logarithm) were able to explain 51.9% of the variance in holistic speaking proficiency scores in the training set. The strongest predictor of speaking proficiency was BNC Trigram Written Frequency Logarithm, which accounted for 35% of the variance in holistic speaking proficiency scores. Speaking samples that had more high-frequency trigrams tended to receive higher scores, suggesting that the use of multiword units that are frequent in a written corpus is an indicator of speaking proficiency. The logged range of content words in a corpus of U.S. subtitles (SUBTLEXus Range CW Logarithm) was the next most important predictor in the model, adding 9.6% of variance explained. This index of range was negatively correlated with speaking proficiency scores, indicating that speaking samples that are judged to be of high proficiency tend to use content words that occur in fewer contexts than the content words used in low proficiency speaking samples. High-proficiency speaking samples also tended to include more academic words, words that were less familiar, and content words that are more frequent than lower proficiency speaking samples, though these last three explained only 1.8%–2.8% of the variance.

Summary of Indices

TAALES was designed as a tool for the automatic analysis of lexical sophistication that can be used by a wide range of educators/researchers (especially those without computer programming skills). This study has demonstrated that the indices included in TAALES are valid measures of lexical sophistication. This study has not only revalidated classic indices of lexical sophistication, but also introduced a number of newly developed indices that are now available for a range of uses in a variety of contexts. These include range indices, bigram indices, and indices that measure academic formulas.

Across the two studies conducted, 55 of the 135 indices investigated were uniquely correlated with holistic proficiency scores in a statistical ($p < .05$) and meaningful ($r > .1$) manner. Additionally, indices from each of the five categories of lexical sophistication investigated (frequency, range, academic lists, psycholinguistic word information, and n-grams) were included in at least one of the multiple regression predictor models. These indices were able to account for 47.5% of the variance in holistic scores of lexical proficiency and 48.7% of the variance in holistic scores of speaking proficiency in the test sets. This suggests that the indices of frequency, range, academic lists, psycholinguistic word information, and n-grams that are included in TAALES are important indicators of lexical proficiency and speaking proficiency.

Previous studies have built a strong case for the importance of word frequency for both receptive (Koda, 2005) and productive (e.g., Crossley, Subtirelu, et al., 2013; Laufer & Nation, 1995) skills, and frequency has had a large impact on language teaching curriculum. The findings of this study suggest that although frequency may indeed be an important indicator of written lexical proficiency and spoken proficiency, range and n-gram indices may be even more important. This would suggest, at least for courses that focus on productive skills, that target vocabulary lists take into account range information. Additionally, the findings of this study indicate that frequent bigrams and trigrams should receive instructional focus.

Frequency indices. With the exception of Brown Frequency CW Logarithm, which played a small role in the speaking proficiency regression model, none of the word frequency measures available in TAALES were selected as predictor variables. Many of these frequency variables were significantly correlated with holistic scores of proficiency (see the Appendix), but were not selected in the regression models indicating that other indices are better predictors of proficiency judgments. Furthermore, the logarithm-transformed frequency of content words based on Brown (1984; Brown Frequency CW Logarithm) correlate positively with holistic speaking proficiency scores, supporting the findings of Crossley et al. (2010, 2014) in that L2 speakers produced more frequent words as a function of time spent learning English. However, these findings generally diverge from other frequency research involving L2 writers (Crossley et al., 2011a; Laufer & Nation, 1995) which demonstrates that more advanced L2 writers produce more infrequent words than beginning writers.

Range indices. An important finding of this study is the strength of range indices to explain human judgments of proficiency. The predictor models in both studies included an index of range. These indices explained 25.2% of holistic lexical proficiency scores and 9.6% of holistic speaking proficiency scores. In each study, range values were negatively correlated with holistic scores, indicating that words that are used in fewer contexts are judged to be more sophisticated than those that occur in more contexts.

N-gram indices. This study provided evidence for the strength of the TAALES n-gram frequency measures. N-gram frequency measures were strong predictors in explaining the variance in lexical proficiency scores (two n-gram indices accounted for 22.3% of the variance) and in explaining the variance in holistic speaking proficiency scores (a single index accounted for 35.2% of the variance).

Academic list indices. Academic list indices had limited success in the two studies. The index ALL AWL RAW was included as a predictor variable in holistic speaking proficiency model (and explained 2.8% of the variance). Almost all other academic list variables, however, were not distributed normally, mostly due to the number of samples that included no instances from the lists. None of the AFL indices were normally distributed in either of the data sets. Additionally, none of the 20 AWL sublist indices were normally distributed in either of the corpora due to the high number of samples that included no AWL sublist words. Although the AFL (and AWL sublist) indices were not important indicators of lexical sophistication in either of the two studies we conducted, it is possible that these indices may be predictive in longer texts.

Word information indices. Word information indices contributed to the variance accounted for by the regression model in both studies. None of the Kuperman et al. (2012) AoA indices were selected for any of the regression models, although the all word and content word indices were significantly correlated with holistic scores of lexical and speaking proficiency. The classic MRC (Coltheart, 1981) concreteness indices and the newer Brysbaert et al. (2013) indices correlated similarly with holistic scores in each study, though none were included in a regression model.

Applications

Second language acquisition. This study has important implications for theories of second language acquisition. The results of the study suggest that as L2 lexical proficiency develops, L2 writers produce a higher proportion of bigrams that are commonly used in spoken language, lexical items that have a narrower range, more frequent trigrams, and lexical items that are less familiar and less meaningful. Furthermore, the results of this study suggest that as L2 speaking proficiency develops, L2 speakers use more high-frequency trigrams, content words that have a narrower range, more academic words, and less familiar and more frequent lexical items.

These findings both support and add to current theories of language acquisition. First, current theories of language development (e.g., Ellis, 2002) suggest that as L2 learners develop, they will use less frequent lexical items. This study presents a new perspective by using spoken unigram frequencies taken from the Brown corpus (500,000 words), which were positively correlated with proficiency scores. In cross-sectional studies (though see the longitudinal studies by Crossley et al., 2010, 2014) this is a novel finding that warrants further examination. Crossley and McNamara (2013), for example, using the same speaking corpus as the present study, found that written frequency scores, as compared to spoken frequency scores, based on a larger corpus (CELEX; 17.6 million words) were negatively correlated with judgments of speaking proficiency. The differences in these two studies suggest that the concept of word frequency in L2 language assessment is complex. However, because the Brown frequencies are based on the much smaller London-Lund corpus, size may also be an important factor when considering the relationship between frequency and speaking proficiency. Clearly, this is an area where more research is needed.

In contrast to previous studies regarding unigram frequency, this study also suggests that as L2 learners develop they tend to use more frequent multiword units (bigrams and trigrams). This study also supports previous studies that have suggested that as L2 learners develop, they produce words that occur in fewer contexts (e.g., Crossley, Subtirelu, et al., 2013). Furthermore, this study supports previous findings that suggest that as L2 learners develop they use words that are less meaningful (Salsbury, Crossley, & McNamara, 2011) and less familiar (Crossley et al., 2011b).

Language assessment. This study also has important implications for L2 language assessment. Perhaps the most salient implication is for the development of models to automatically assess language proficiency. Although word-level (unigram) frequency has had an impact on L2 language assessment models (e.g., Attali & Burstein, 2006), models of language assessment can clearly benefit from the inclusion of range, bigram, and trigram indices. This study has demonstrated that proficient L2 users produce unigrams that are used in fewer contexts than the unigrams produced by less proficient L2 users. Such indices could be used in future models to improve accuracy and potential feedback provided to users. Furthermore, this study has demonstrated that more proficient users also use bigrams and trigrams that are more frequent than less proficient users and that bigram and trigram frequency indices are more predictive of language proficiency than their unigram counterparts. This indicates that future language assessment models could benefit from the inclusion of bigram and trigram frequencies in addition to unigram frequencies.

This study also has important implications for traditional standardized L2 language assessments, institution-level L2 language assessments, and classroom-level L2 assessments that do not rely on automatic models. The findings indicate that assessment measures could be refined by taking into account not only the importance of word-level frequency, but also the importance of word range and n-grams as significant factors of L2 proficiency. Tests that assess lexical knowledge (e.g., frequency based recognition and production tests; Meara, 1992; Nation, 1990, 2001; Schmitt, Schmitt, & Clapham, 2001) based on single word frequency counts could be improved through the inclusion of range and bigram frequency considerations.

Classroom pedagogy. This study has important implications for classroom pedagogy. First and foremost, the target lexicon for productive language instruction may benefit from selection that is based on additional factors beyond word-level frequency counts. This study has demonstrated that the production of frequent bigrams and trigrams is more predictive of L2 lexical and speaking proficiency, indicating that it may be more important for language learners to focus on how words are frequently combined than on learning infrequent words and that using more frequent words may be preferred in spoken contexts. Thus, this finding suggests that L2 productive language classrooms should, in addition to single words, focus on multiword units that are chosen based on frequency.

This study also suggests that when word-level vocabulary is introduced in L2 classrooms, range needs to be considered. This study has demonstrated that range is a stronger predictor of judgments of both lexical and spoken proficiency than word-level frequency counts, suggesting that the construct of lexical sophistication is much more complex than once posited. It is important for L2 students to not only learn words that are less frequent, but also learn words that are more context-specific (i.e., have a narrower range). Indeed, with regard to spoken frequency lists and spoken tasks, it may be beneficial for learners to learn words that are generally frequent in spoken corpora but have a restricted range (i.e., a lower number of unique texts in which the words occur). In some contexts this may include the AWL, but the findings of this study indicate that the use of AWL entries plays but a small role in judgments of L2 speaking proficiency. Although the AWL did indeed play a small role in the judgments of L2 speaking proficiency, overall academic lists did not play a particularly important role in predicting the judgments of L2 lexical or speaking proficiency. Both the AWL and (to a lesser degree) the AFL have been shown to be important for academic receptive skills (e.g., Coxhead, 2000; Simpson-Vlach & Ellis, 2010). This study indicates that they may not be useful targets in less formal productive contexts such as TOEFL independent speaking tasks and early undergraduate free writing (lexical proficiency corpus), a finding that is not entirely surprising given the more formal nature of the corpora that the AWL and the AFL are derived from.

CONCLUSION

This study introduced TAALES, a freely available, easy-to-use tool that includes 135 classic and newly developed indices of lexical sophistication. Indices of frequency, range, academic lists, psycholinguistic word information, and n-grams were validated in studies that examined holistic scores of lexical proficiency and speaking proficiency. These resources provide additional support for researchers interested in L2 learning, L2 assessment, and L2 classroom instruction.

Limitations

Although TAALES has proven to be a viable method from which to examine lexical production, it is by no means inclusive. None of the indices included in TAALES directly measures lexical accuracy or contextual appropriateness, nor do they measure word polysemy or hypernymy. Additionally, although some of the n-gram indices may indirectly measure collocation knowledge, this article has not explored the validity of TAALES indices to do so, though this may hold promise for future research. Furthermore, although this study has explored the amount of variance in written lexical proficiency and spoken proficiency explained by lexical sophistication indices in the areas of word frequency, word range, n-grams, academic lists, and psycholinguistic word information, it did not directly assess judgments of writing quality. Finally, due to the limited size of some of the databases used by TAALES (e.g., Brown, 1984, frequency values), not all words in a target text are necessarily used to calculate lexical sophistication scores, which may result in an underuse of source data.

Future Directions

Future research should address the limitations outlined above. Perhaps the most straightforward research task will be to assess whether and to what degree indices related to lexical frequency, range, n-gram frequency, academic vocabulary, and psycholinguistic word properties explain the variance in holistic scores of writing quality. Additionally, the degree to which n-gram indices approximate collocational knowledge should be explored. Finally, because TAALES is intended to be accessible to a wide range of educators and researchers, usability studies of the program should be conducted in a variety of environments to determine its usability and application in second language research, lexical assessment, and the second language classroom.

ACKNOWLEDGMENTS

We would like to thank all who helped us test the functionality and usability of TAALES. We would especially like to thank Sarah Goodwin, Jessica Kyle, Cornelia

Lahmann, Cindy Murphy, Geoffrey Pinchbeck, and Stephen Skalicky. We would also like to thank Danielle McNamara for her valuable insights during the early phases of the development of TAALES. Finally, we would like to extend our thanks to Tom Salsbury for granting us access to the lexical proficiency corpus, to Scott Jarvis for assisting with the creation of the lexical proficiency rubric, and to all of the raters who provided the lexical proficiency scores.

THE AUTHORS

Kristopher Kyle is a doctoral student in the Department of Applied Linguistics and ESL at Georgia State University. His research interests include second language writing and speaking, assessment, and second language acquisition. He is especially interested in applying natural language processing and corpora to the exploration of these areas.

Scott A. 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 the development and application of computational tools in second language learning and text comprehensibility.

REFERENCES

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jtla.org

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at . . . : Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405. doi:10.1093/applin/25.3.371

BNC Consortium. (2007), British National Corpus, version 3 (BNC XML ed.). Retrieved from http://www.natcorp.ox.ac.uk

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instrumentation & Computers, 16*, 502–532. doi:10.3758/BF03200836

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990. doi:10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods.* Advance online publication. doi:10.3758/s13428-013-0403-5

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology, 11*(3), 38–63. Retrieved from http://llt.msu.edu

Cobb, T. (2013). *Web Vocabprofile.* Retrieved from http://www.lextutor.ca/vp

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology Section A, 33*, 497–505. doi:10.1080/14640748108400805

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213–238. doi:10.2307/3587951

Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, *45*, 355–362. doi:10.5054/tq.2011.254528

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 214–219) Menlo Park, CA: AAAI Press.

Crossley, S. A., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*(3), 250–270. doi:10.1080/15434303.2014.926905

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*, 965–981. doi:10.1016/j.system.2013.08.002

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*, 475–493. doi:10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, *35*, 115–135. doi:10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*, 171–192.

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605. doi:10.1111/j.1467-9922.2010.00568.x

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). Predicting lexical proficiency in language learners using computational indices. *Language Testing*, *28*, 561–580. doi:10.1177/0265532210378031

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, *45*, 182–193. doi:10.5054/tq.2010.244019

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning: What predicts early lexical production? *Studies in Second Language Acquisition*, *35*, 727–755. doi:10.1017/S0272263113000375

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*, 282–311. doi:10.1177/0741088311410188

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2*, 151–177.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188. doi:10.1017/S0272263102002024

Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*, 414–420. doi:10.2307/3587446

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*, 193–202. doi:10.3758/BF03195564

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403–437. doi:10.1075/ijcl.13.4.02gri

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing.* Ann Arbor: University of Michigan Press.

Jarvis, S., & Crossley, S. A. (2012). *Approaching language transfer through text classification: Explorations in the detection-based approach.* Bristol, England: Multilingual Matters.

Koda, K. (2005). *Insights into second language reading.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139524841

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44,* 978–990. doi:10.3758/s13428-012-0210-4

Kyle, K., Crossley, S. A., Dai, J., & McNamara, D. S. (2013, June). *Native language identification: A key ngrams approach.* Paper presented at the 8th Workshop on Innovative Use of NLP for Building Educational Applications, Atlanta, GA.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*(2), 21–33. doi:10.1177/003368829402500202

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16,* 307–322. doi:10.1093/applin/16.3.307

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42,* 381–392. doi:10.3758/BRM.42.2.381

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57–86. doi:10.1177/0741088309351547

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188–205). Hershey, PA: IGI Global. doi:10.4018/978-1-60960-741-8

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge, England: Cambridge University Press. doi:http://dx.doi.org/10.1017/CBO9780511894664

Meara, P. (1992). *EFL vocabulary tests.* Swansea, Wales: Centre for Applied Language Studies.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge, England: Cambridge University Press.

Meara, P. (2005a). Designing vocabulary tests for English, Spanish and other languages. In C. Butler, S. Christopher, M. Á. Gómez González, & S. M. Doval-Suárez (Eds.), *The dynamics of language use* (pp. 271–285). Amsterdam, the Netherlands: John Benjamins.

Meara, P. (2005b). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26,* 32–47. doi:10.1093/applin/amh037

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of TESL student performance. *System, 32,* 75–87. doi:10.1016/j.system.2003.05.001

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Boston, MA: Heinle and Heinle. doi:10.1016/0346-251X(94)90065-5

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139524759

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count.* Retrieved from http://homepage.psy.utexas.edu/HomePage/Faculty/Pennebaker/Reprints/LIWC2007_OperatorManual.pdf

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27,* 343–360. doi:10.1177/0267658310395851

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing, 18,* 55–88. doi:10.1191/026553201668475857

Simpson, R., Briggs, S., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English.* Ann Arbor: Regents of the University of Michigan.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31,* 487–512. doi:10.1093/applin/amp058

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation.* Lund, Sweden: Gleerup.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York, NY: Teachers College, Columbia University.

Toglia, M. P., & Battig, W. R. (1978). *Handbook of semantic word norms.* New York, NY: Lawrence Erlbaum.

Witten, I. A., Frank, E., & Hall, M. A. (2011). *Data mining.* San Francisco, CA: Elsevier.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3,* 215–222.

APPENDIX
Summary of All Indices of Lexical Sophistication Investigated

Variable	Correlation with lexical proficiency	Correlation with speaking proficiency	Included in lexical proficiency regression	Included in speaking proficiency regression
Frequency Indices				
BNC Written Freq AW	0.115	0.054		
BNC Written Freq AW Log	-0.140*	0.040	✓	
BNC Written Freq CW	-0.140*	-0.089	✓	
BNC Written Freq CW Log	-0.270**	-0.050	✓	
BNC Written Freq FW	0.136*	0.055		
BNC Written Freq FW Log	0.173**	0.098		
BNC Spoken Freq AW	0.177**	0.070	✓	
BNC Spoken Freq AW Log	-0.026	0.031		
BNC Spoken Freq CW	-0.085	-0.085		
BNC Spoken Freq CW Log	-0.124	-0.033		
BNC Spoken Freq FW	0.194**	0.084	✓	
BNC Spoken Freq FW Log	0.166**	0.101	✓	
Brown Freq AW	0.228**	0.172**	✓	✓
Brown Freq AW Log	0.100	0.200**		✓
Brown Freq CW	0.065	0.176**		✓
Brown Freq CW Log	-0.013	0.146*		✓
Brown Freq FW	0.191**	0.107	✓	
Brown Freq FW Log	0.156*	0.102		
KF Freq AW	0.140*	0.079		
KF Freq AW Log	-0.041	0.147*		✓
KF Freq CW	-0.114	0.066		
KF Freq CW Log	-0.220**	0.054	✓	
KF Freq FW	0.130*	0.040		
KF Freq FW Log	0.166**	0.074		
SUBTLEXus Freq AW	0.132*	0.076	✓	
SUBTLEXus Freq AW Log	-0.099	-0.043		
SUBTLEXus Freq CW	-0.170**	-0.085	✓	
SUBTLEXus Freq CW Log	-0.210**	-0.121		
SUBTLEXus Freq FW	0.137*	0.103	✓	
SUBTLEXus Freq FW Log	0.139*	0.091		
TL Freq AW	0.191**	0.097	✓	
TL Freq AW Log	-0.026	0.126*		✓
TL Freq CW	-0.078	0.095		
TL Freq CW Log	-0.160**	0.031	✓	
TL Freq FW	0.181**	0.043	✓	
TL Freq FW Log	0.167**	0.075	✓	

Appendix *(Continued)*

Variable	Correlation with lexical proficiency	Correlation with speaking proficiency	Included in lexical proficiency regression	Included in speaking proficiency regression
Range Indices				
BNC Written Range AW	-0.200**	-0.014	✓	
BNC Written Range CW	-0.250**	-0.028		
BNC Written Range FW	0.161*	-0.037	✓	
BNC Spoken Range AW	-0.024	0.037		
BNC Spoken Range CW	-0.054	0.034		
BNC Spoken Range FW	0.080	Non-Normal		
KF Ncats AW	-0.260**	-0.03		
KF Ncats CW	-0.290**	-0.041	✓	
KF Ncats FW	Non-Normal	Non-Normal		
KF Nsamp AW	-0.004	0.136*		✓
KF Nsamp CW	-0.140*	0.089		
KF Nsamp FW	0.154*	0.016	✓	
SUBTLEXus Range AW	-0.116	-0.045		
SUBTLEXus Range AW Log	-0.190**	-0.15*		
SUBTLEXus Range CW	-0.140*	-0.054		
SUBTLEXus Range CW Log	-0.220**	-0.160**	✓	✓
SUBTLEXus Range FW	Non-Normal	Non-Normal		
SUBTLEXus Range FW Log	Non-Normal	Non-Normal		
N-Grams				
Bigram Type Count	0.193**	0.737**		
Trigram Type Count	0.100	0.724**		
BNC Written Bigram Freq Log	0.155*	0.672**		
BNC Written Bigram Freq Normed (bi)	Non-Normal	-0.013		
BNC Written Bigram Freq Normed (bi) Log	0.073	0.248**		✓
BNC Written Bigram Freq Normed (word)	Non-Normal	0.024		
BNC Written Bigram Freq Normed (word) Log	0.233**	0.248**		✓
BNC Written Trigram Freq Log	0.296**	0.593**		✓
BNC Written Trigram Freq Normed (tri)	Non-Normal	0.045		
BNC Written Trigram Freq Normed (tri) Log	0.061	0.045		

Appendix *(Continued)*

Variable	Correlation with lexical proficiency	Correlation with speaking proficiency	Included in lexical proficiency regression	Included in speaking proficiency regression
BNC Written Trigram Freq Normed (word)	Non-Normal	0.184**		✓
BNC Written Trigram Freq Normed (word) Log	0.352**	0.250**	✓	✓
BNC Spoken Bigram Freq Log	0.185**	0.643**	✓	
BNC Spoken Bigram Normed (bi) Freq	0.034	-0.031		
BNC Spoken Bigram Normed (bi) Freq Log	0.135*	0.181**	✓	✓
BNC Spoken Bigram Normed (word) Freq Log	0.145*	0.007	✓	
BNC Spoken Bigram Normed (word) Freq	0.274**	0.166**	✓	✓
BNC Spoken Trigram Freq Log	0.309**	0.548**	✓	
BNC Spoken Trigram Normed (tri) Freq	Non-Normal	0.068		
BNC Spoken Trigram Normed (tri) Freq Log	0.023	0.035		
BNC Spoken Trigram Normed (word) Freq Log	Non-Normal	0.133*		✓
BNC Spoken Trigram Normed (word) Freq	0.361**	0.184**		✓
BNC Written Bigram Proportion	0.372**	0.212**		✓
BNC Spoken Bigram Proportion	0.409**	0.166**	✓	
BNC Written Trigram Proportion	0.395**	0.316**		✓
BNC Spoken Trigram Proportion	0.400**	0.259**	✓	✓
Academic List Indices				
All AFL Raw	Non-Normal	Non-Normal		
All AFL Normed	Non-Normal	Non-Normal		
Core AFL Raw	Non-Normal	Non-Normal		
Core AFL Raw	Non-Normal	Non-Normal		
Spoken AFL Raw	Non-Normal	Non-Normal		
Spoken AFL Normed	Non-Normal	Non-Normal		
Written AFL Raw	Non-Normal	Non-Normal		
Written AFL Normed	Non-Normal	Non-Normal		
All AWL Raw	Non-Normal	0.341**		✓
All AWL Normed	Non-Normal	0.094		
EACH AWL Sublist, Raw and Normed	Non-Normal	Non-Normal		

Appendix *(Continued)*

Variable	Correlation with lexical proficiency	Correlation with speaking proficiency	Included in lexical proficiency regression	Included in speaking proficiency regression
Word Information Indices				
Kuperman AoA AW	0.163*	0.197**		
Kuperman AoA AW Log	0.171**	0.201**		
Kuperman AoA CW	0.171**	0.228**		
Kuperman AoA CW Log	0.180**	0.235**	✓	✓
Kuperman AoA FW	0.079	0.031		
Kuperman AoA FW Log	0.086	0.034		
Brysbaert Concreteness Unigram AW	-0.270**	-0.300**		✓
Brysbaert Concreteness Unigram CW	-0.170**	-0.320**		
Brysbaert Concreteness FW	-0.250**	-0.039	✓	
Brysbaert Concreteness Bigrams AW	Non-Normal	Non-Normal		
Brysbaert Concreteness Combined AW	-0.290**	-0.300**	✓	
Brysbaert Concreteness Combined CW	-0.180**	-0.320**	✓	✓
Brysbaert Concreteness Combined FW	-0.210**	-0.051		
MRC Concreteness AW	-0.290**	-0.330**		
MRC Concreteness CW	-0.170**	-0.290**		
MRC Concreteness FW	-0.180**	-0.073		
MRC Familiarity AW	-0.200**	-0.230**	✓	✓
MRC Familiarity CW	-0.330**	-0.350**	✓	✓
MRC Familiarity FW	0.079	0.024		
MRC Imageability AW	-0.350**	-0.360**	✓	✓
MRC Imageability CW	-0.230**	-0.300**	✓	✓
MRC Imageability FW	-0.220**	-0.084		
MRC Meaningfulness AW	-0.380**	-0.330**	✓	✓
MRC Meaningfulness CW	-0.300**	-0.250**		
MRC Meaningfulness FW	-0.230**	-0.081	✓	

*p < .05; **p < .01.

Note. AW = all words; CW = content words; FW = function words; Freq = frequency; Log = logarithm.