The Modern Language Journal

Lexical Sophistication as a Multidimensional Phenomenon: Relations to Second Language Lexical Proficiency, Development, and Writing Quality

MINKYUNG KIM,[1] SCOTT A. CROSSLEY,[1] and KRISTOPHER KYLE[2]
[1] *Georgia State University, Applied Linguistics and ESL, PO Box 4099, Atlanta, GA, 30302–4099*
Email: mkim89@gsu.edu, scrossley@gsu.edu
[2] *University of Hawai'i at Manoa, Second Language Studies, 587 Moore Hall, 1890 East West Road, Honolulu, HI, 96822 Email: kkyle@hawaii.edu*

This study conceptualizes lexical sophistication as a multidimensional phenomenon by reducing numerous lexical features of lexical sophistication into 12 aggregated components (i.e., dimensions) via a principal component analysis approach. These components were then used to predict second language (L2) writing proficiency levels, holistic lexical proficiency scores, and longitudinal lexical growth. The results from regression analyses indicated that 5 lexical components (i.e., bigram and trigram strength of directional association, content word properties, bigram mutual information, bigram and trigram proportions, and word specificity) explained 16.1% and 31.0% of the variance of L2 writing proficiency and lexical proficiency, respectively. Two additional components (i.e., word acquisition properties and content word frequency) explained an additional 8.5% of the variance of L2 writing proficiency. Six lexical components (i.e., bigram and trigram proportions, word acquisition properties, content word frequency, bigram frequency and range, content word properties, and function word frequency and range) showed significant developmental trends in L2 beginning learners over a year-long period. These findings provide information about the multidimensional nature of lexical sophistication by expanding its scope beyond frequency and toward other primary dimensions that include various lexical and phrasal features such as concreteness, orthographic density, hypernymy, and *n*-gram frequency and association strength.

Keywords: lexical sophistication; vocabulary; proficiency; corpus linguistics

OVER THE PAST TWO DECADES, A growing body of literature has investigated second language (L2) learners' lexical proficiency (e.g., Crossley et al., 2011a, 2011b; Daller, Milton, & Treffers–Daller, 2007; Meara, 1996; Nation, 1990; Read, 2000). Although an overall definition of lexical proficiency has yet to be agreed upon, it is generally accepted that lexical proficiency, like other constructs of language proficiency, includes both lexical knowledge and lexical use in context along with metacognitive strategies for lexical use (Chapelle, 1994; Read, 2000). A substantial body of research has investigated learners' productive vocabulary under the assumption that learners' vocabulary use may reflect learners' lexical knowledge, be closely aligned with general language proficiency levels, and indicate stages of vocabulary acquisition (Crossley et al., 2011a, 2011b; Daller & Xue, 2007; Jarvis, 2013; Lu, 2012; Saito et al., 2016).

Although various terms have been used in defining and measuring L2 learners' productive vocabulary, *lexical richness* (i.e., the overall

The Modern Language Journal, 00, 0, (2018)
DOI: 10.1111/modl.12447
0026-7902/18/1–22 $1.50/0

quality of vocabulary found in a language sample; Daller et al., 2007; Milton, 2009; Read, 2000) is perhaps one of the most widely used terms. Lexical richness is operationalized into various measures—including, but not limited to, lexical sophistication and lexical diversity (Jarvis, 2013). Lexical sophistication refers to the learner's use of sophisticated and advanced words while lexical diversity refers to the overall range of unique words used. In terms of lexical sophistication, the definition of sophisticated, difficult, or advanced words has not been agreed upon (Daller et al., 2007). Sophisticated words were originally associated with low-frequency words (Laufer & Nation, 1995), but word sophistication may be defined beyond frequency (Kyle & Crossley, 2015, 2016). For instance, recent investigations using computational indices have expanded the scope of lexical sophistication by including various lexical features such as range, *n*-gram frequency, academic words, hypernymy, polysemy, and psycholinguistic word information related to concreteness, imageability, meaningfulness, familiarity, and age of acquisition (Crossley, Clevinger, & Kim, 2014; Crossley et al., 2011a, 2011b; Crossley, Salsbury, & McNamara, 2009, 2010, 2015; Crossley, Subtirelu, & Salsbury, 2013; Kyle & Crossley, 2015; Saito et al., 2016). Accordingly, lexical sophistication is likely multidimensional (Lu, 2012), and conceptually distinct lexical features may measure similar constructs (e.g., word frequency and word familiarity might both tap into word exposure). In addition, there are many possible methods to calculate a lexical feature, and each method may measure distinct properties of that feature. For instance, the feature *lexical frequency* may consist of a number of 'micro-features' such as content word frequency, function word frequency, spoken frequency, and/or written frequency. Thus, we presume that there are 'macro-features' or dimensions of lexical sophistication that are a combination of similar features and/or co-occurring micro-features. However, there have been few attempts at investigating the multidimensional nature of lexical sophistication.

As an exploratory study, the main purpose of the current study is to conceptualize lexical sophistication as a multidimensional phenomenon through computational evaluation of lexical and phrasal features as found in learner corpora. To this end, we first develop macro-features (i.e., components or dimensions) of lexical sophistication based on statistical co-occurrences of lexical micro-features using a principle component analysis (PCA). The lexical micro-features for this study are reported by the Tool for the Automatic Analysis of Lexical Sophistication 2.0 (TAALES 2.0; Kyle & Crossley, 2015). Our secondary interest is in assessing the validity of the developed macro-features using corpora related to L2 proficiency. Thus, we examine three L2 corpora using the developed macro-features to assess three different language proficiency criteria: L2 writing proficiency levels, holistic lexical proficiency scores, and L2 longitudinal growth. Overall, we examine evidence of interaction among micro-features, identify primary dimensions of lexical sophistication, and provide evidence for the importance of large-scale lexical sophistication dimensions in describing lexical and writing proficiency scores and lexical development in L2 learners. Thus, this study is guided by four research questions (RQs):

RQ1. How can lexical micro-features be grouped into lexical macro-features?
RQ2. Can lexical macro-features be used to predict L2 writing proficiency?
RQ3. Can lexical macro-features be used to predict lexical proficiency in an L2 writing corpus?
RQ4. Can lexical macro-features be used to track L2 beginning learners' longitudinal development of lexical sophistication using a one-year corpus of L2 spoken samples?

LEXICAL SOPHISTICATION

A traditional definition of lexical sophistication is the proportion of relatively advanced words in the learner's sample (Read, 2000). Sophisticated or advanced words are generally conceptualized as low-frequency words (Laufer & Nation, 1995). This conceptualization has been warranted by L2 vocabulary acquisition literature that explored the links between the knowledge and use of less frequent words and general language proficiency (Daller et al., 2007; Meara, 1996; Milton, 2009) as well as L2 assessment literature in evaluating the relationships between lexical sophistication and human judgments of L2 writing, speaking, and lexical proficiency (Crossley et al., 2011a, 2011b; Kyle & Crossley, 2015, 2016). In general, these studies report that greater lexical/language proficiency is associated with the use of less frequent words.

Beyond frequency, there has been little agreement on how sophisticated words should be defined (Daller & Xue, 2007; Milton, 2009; Read, 2000). For instance, sophisticated words have been categorized as words that are less concrete, less imageable, and less familiar (Salsbury,

Crossley, & McNamara, 2011; Saito et al., 2016), words that are contextually less diverse (McDonald & Shillcock, 2001), words with fewer phonological and orthographical neighbors (Balota et al., 2007), words that are more specific (Fellbaum, 1998), words that elicit slower response times in lexical decision tasks (Balota et al., 2007), and words that are widely used in academic contexts (Coxhead, 2000). Furthermore, sophisticated lexical items can be extended into multiword units based on the notion that knowing how two or more words combine into multiword units is an essential element of lexical knowledge (Sinclair, 1991).

The various categorizations of sophisticated words are based on differences in disciplines (e.g., applied linguistics, psycholinguistics, and computational linguistics) and theoretical frameworks. For instance, usage-based approaches assume that frequency of occurrence has a substantial influence on language acquisition. The more frequently people experience a linguistic representation, the stronger it is entrenched (i.e., stored) in their memory and, thus, the earlier it is acquired (Ellis, 2002). In addition, usage-based theories presume that lexical processes involve associative learning (i.e., establishment of connections), abstraction (i.e., categorizing instances into schemata), automatization (i.e., use of language without cognitive efforts), and developing representations of form and meaning (Ellis, 2002; Ellis & Larsen–Freeman, 2009; Gries & Ellis, 2015; Langacker, 2007). Importantly, these form–function pairings include not just individual words, but also strings of words used as chunks (i.e., collocations; Goldberg, 2006). The view that lexical units are conceptualized at both word and phrase levels is a phraseological perspective, which suggests that people comprehend, process, and use collocations in a manner similar to their use of single lexical items (Hoey, 2005; Sinclair, 1991; see Evert, 2008, for detailed discussion of collocations).

Psycholinguistic perspectives presume that various lexical features (e.g., concreteness and orthographic density) have considerable influences on word recognition and processing (e.g., Balota et al., 2007; Coltheart, 1981). For instance, response latencies and accuracies may be indications of lexical sophistication such that words that elicit longer response times and fewer accuracies may be more sophisticated. In this respect, more proficient speakers likely use words that are cognitively more difficult to recognize and process. On the other hand, from a computational linguistic perspective, lexical sophistication of words can be computed based on co-occurrence patterns over a large corpus of text such that more sophisticated words may co-occur with a narrow set of related words (Dascalu et al., 2016).

COMPUTATIONAL ANALYSES OF LEXICAL SOPHISTICATION

Computational analyses of the words and phrases produced by L2 learners and L2 speakers have provided a wealth of knowledge about lexical sophistication (e.g., Crossley et al., 2011a, 2011b; Kyle & Crossley, 2015; Lu, 2012). We provide an overview of these findings in terms of the various domains of lexical sophistication.

Word Frequency and Range

Frequency information has been extensively examined under the notion that words that are frequently heard and read by L2 learners tend to be learned earlier than words that are infrequently heard and read (Ellis, 2002; Gries & Ellis, 2015). Word frequency is predictive of human ratings of L2 lexical proficiency (Crossley et al., 2011a, 2011b; Kyle & Crossley, 2015), L2 writing proficiency (Crossley & McNamara, 2012), and L2 speaking proficiency (Kyle & Crossley, 2015), such that more proficient L2 learners tend to produce less frequent words. Word range, as compared to frequency, examines how widely a given word occurs across documents within a corpus (Kyle & Crossley, 2015) and addresses concerns that frequency can be conflated by the distributional patterns of words across a reference corpus (Crossley et al., 2013). Like word frequency, word range is predictive of L2 lexical proficiency and L2 speaking proficiency such that more proficient L2 learners tend to use words that occur in fewer contexts than less proficient L2 learners (Kyle & Crossley, 2015).

Contextual Distinctiveness

Researchers have examined lexical and semantic relationships between a word and the contexts in which the word occurs, arguing that word co-occurrence is a strong predictor of word learning and processing (i.e., a word's contextual distinctiveness). For instance, McDonald and Shillcock (2001) showed that words that appear in relatively constrained linguistic contexts (i.e., have high distinctiveness) have longer lexical decision latencies than words whose contexts are unconstrained (i.e., have low distinctiveness). Crossley et al. (2013) examined the probability

of both frequency and contextual distinctiveness indices to predict noun and verb production in L2 speech. They reported that word frequency was the strongest classifier of whether a noun was produced or not produced in beginning L2 oral discourse, while contextual diversity was the strongest classifier of verb production.

Psycholinguistic Properties of Words

A number of psychological properties of words based on subjective human ratings have been extensively examined, including meaningfulness, concreteness, familiarity, imageability, and age of acquisition (AoA; Brysbaert, Warriner, & Kuperman, 2014; Coltheart, 1981; Kuperman, Stadthagen–Gonzales, & Brysbaert, 2012; Salsbury et al., 2011). Word meaningfulness refers to how likely words are to activate other words. Words with high meaningfulness scores (e.g., *cup)* evoke many word associations, while words with low meaningfulness scores (e.g., *astuteness)* evoke fewer word associations.[1] The use of words with low meaningfulness scores is a predictor of high L2 language proficiency levels (Salsbury et al., 2011) and L2 lexical proficiency (Kyle & Crossley, 2015). Word concreteness is related to how perceptible and tangible the mental representation of a word is and is quite similar to word imageability, which is related to how easily a word evokes a mental and sensory image. Concrete words (e.g., *mountains*) tend to be learned earlier than abstract words (e.g., *honesty*), both in first languages (L1s; Paivio, 1986) and L2s (Kaushanskaya, & Rechtzigel, 2012).[2] Word familiarity is related to how commonly words are experienced, and closely associated with word frequency (i.e., text-based measures of occurrences of words) and word exposure (Schmitt & Meara, 1997). Word familiarity is predictive of L2 speaking proficiency and L2 lexical proficiency, indicating that more advanced L2 learners tend to use less familiar words (Kyle & Crossley, 2015). Last, words with lower AoA (e.g., *lunch*), which is based on human judgments of the age at which a word is learned (Kuperman et al., 2012), are processed more quickly than words with higher AoA scores (e.g., *luminosity*).[3]

Phonological, Orthographical, and Phonographic Neighbors

Researchers (e.g., Balota et al., 2007) have also identified phonological neighborhood (i.e., phonological neighbors differing by one phoneme; e.g., *gate* and *get*) and orthographical neighborhood (i.e., orthographic neighbors differing by one letter; e.g., *cough* and *dough*) as important features of word learning and lexical processing. Previous research has found that an increase in the number of neighbors has facilitating effects on word processing both in L1s (Forster & Shen, 1996) and L2s (van Heuven, Dijkstra, & Grainger, 1998), such that words with orthographically and phonologically dense neighborhoods (e.g., *get*) are considered less sophisticated than words with orthographically and phonologically sparse neighborhoods (e.g., *responsibility*). Adelman and Brown (2007) further found processing effects for phonographic neighborhood (i.e., phonographic neighbors differing in one letter and one phoneme; e.g., *stone* and *stove*) on word naming latencies, such that words with greater neighborhoods were processed more quickly.

Word Recognition

Researchers are also interested in investigating lexical recognition using lexical decision and naming tasks (Balota et al., 2007; Zorzi, Houghton, & Butterworth, 1998). In lexical decision tasks, participants see a string of letters, and decide whether the given string of letters is a word or a nonword. In naming tasks, participants are presented with a word or a nonword and are asked to read the word aloud. The behavior results of these tasks are based on how quickly and accurately participants react to the word or nonword stimuli. These word recognition norms likely reflect the degree of difficulty of a word for L1 speakers based on on-line processing information for the given word. For example, words that elicit longer response times and less accurate responses for L1 speakers (e.g., 1009.94ms for the word *expatriate* in lexical decision tasks) are considered more sophisticated than words that elicit shorter response times and more accurate responses (e.g., 540.55ms for the word *people* in lexical decision tasks). A recent investigation using native speakers' word recognition norms (Berger, Crossley, & Kyle, 2017) found that higher proficiency L2 learners tended to use words that elicited longer response times and less accurate responses in L1 speakers, and that lexical decision latencies explained more variance in L2 lexical proficiency than word frequency.

Age of Exposure

Age of Exposure (AoE) indices reflect the age/grade level of a learner and are based on

computational models that estimate a word's difficulty based on co-occurrence data and a word's links to relevant semantic concepts within large corpora. Using Latent Dirichlet Allocation (LDA) (i.e., computational modeling techniques used to infer underlying topics through a generative probabilistic process), Dascalu et al. (2016) developed measures of AoE values for the words in the Touchstone Applied Science Associates (TASA) corpus which contains 13 grade-level textbooks in the United States (Landauer, Foltz, & Laham, 1998). Their findings indicated that AoE indices were strongly related to human ratings of age of acquisition, word frequency, entropy, and human lexical response latencies.

Hypernymy

Hypernymy reports hierarchical associations between hypernyms (i.e., general words such as *vehicle*) and hyponyms (i.e., specific words such as *car*). Hypernymy represents the degree of word specificity in a semantic hierarchy. Research has shown mixed findings in L2 learners' use of hypernyms. For instance, research has reported that L2 learners are more likely to use words with wide coverage (e.g., *on*) than words with specific coverage (e.g., *onto* and *upon*; Ijaz, 1986; Levenston & Blum, 1977) and that more proficient L2 writers tend to use more specific words than less proficient L2 writers (Guo, Crossley, & McNamara, 2013; Kyle & Crossley, 2016). Other research found that hypernymic relations in L2 learners' output become less specific as a function of time spent studying English (Crossley et al., 2009) and that word hypernymy values are predictive of L2 lexical proficiency such that more advanced L2 learners tend to use less specific words (Crossley et al., 2011a).

Polysemy

Polysemy reports the number of senses associated with a word. Polysemous words have more senses and are thus more ambiguous. Polysemous words are also generally more frequent. Research has shown that as proficiency levels increase, L2 learners improve their word sense knowledge (Schmitt, 1998) and are able to extend the core meanings of polysemous words (Crossley et al., 2010).

Bigram and Trigram Frequency and Range

N-grams (i.e., contiguous sequences of *n* words, such as *of the* and *in the*) are considered important in learning English because *n*-grams occur in over 80% of the words produced in spoken English (Altenberg, 1998). Researchers have examined bigram and trigram frequencies, ranges, and proportions (i.e., the proportion of *n*-grams in a given sample that are also found in a reference corpus) in relation to human ratings of language proficiency (Kyle & Crossley, 2015). Kyle and Crossley (2015) found that *n*-gram frequencies and proportions were predictive of human ratings of L2 lexical proficiency, indicating that L2 language samples rated as having higher lexical proficiency contained more frequent trigrams and used proportionally more bigrams that also occur in the reference corpus. They also found that trigram frequency was the strongest predictor of L2 speaking proficiency, indicating that speaking samples that contained more frequent trigrams were likely to receive higher scores.

Bigram and Trigram Association Strength

Researchers have also examined bigram and trigram association strength measures (i.e., the degree to which a string of words is attached to one another) under the assumption that language learning can be explained based on the probabilities of occurrence of various representations including form–meaning mappings and adjacent words (i.e., associative learning; Gries & Ellis, 2015). Two well-established association measures are Mutual Information (MI) and *t*-score (Bestgen & Granger, 2014). MI highlights *n*-grams made up of low-frequency words (e.g., *exultant triumph*), while *t*-score highlights *n*-grams composed of high-frequency words (e.g., *big win*). Another approach is to compute directional association strength scores. A well-established directional measure is ΔP (Delta P; i.e., the one-way dependency statistic; Allan, 1980). Directionality is considered in computing ΔP because association strengths between two words are not symmetrical (Gries, 2013). For example, in the combination of *artificial* and *intelligence*, the probability of the word *intelligence* occurring at the right given the occurrence of *artificial* at the left is higher (i.e., high ΔP score for *artificial intelligence*) than the probability of *artificial* occurring to the right of *intelligence* because *intelligence* is often followed by words other than *artificial*, such as *agencies* or *sources*. Research has indicated that L2 learners tend to underuse *n*-grams with high MI scores (i.e., *n*-grams composed of lower frequency words) and overuse *n*-grams with high *t*-scores (i.e., *n*-grams composed of higher frequency words; Durrant & Schmitt, 2009).

Academic Word Lists and Formulas

Academic word and phrase lists have been developed to explore the acquisition of academic language (Coxhead, 2000). Two of the most influential academic word lists in L2 research are the Academic Word Lists (AWL; Coxhead, 2000) and the Academic Formulas List (AFL; Simpson–Vlach & Ellis, 2010). The AWL contains academic words identified as frequently recurring in a written academic corpus, while the AFL contains formulaic sequences identified as frequently recurrent patterns in academic written and spoken corpora. Research has shown that using academic words is indicative of higher quality academic writing (Douglas, 2013) and higher L2 speaking proficiency levels (Kyle & Crossley, 2015).

Current Study

Lexical sophistication has been widely studied across disciplines, including vocabulary acquisition research, writing research, language testing, psycholinguistics, and computational linguistics. While research supports the notion that various features of lexical sophistication are interconnected, relatively less is known about whether interconnected lexical features can be aggregated into larger dimensions and what these dimensions would tell us about lexical sophistication. In short, while it is assumed that lexical sophistication is a multidimensional phenomenon, there is scant empirical evidence to support this notion. Thus, the current study develops a large-grain understanding of lexical sophistication by aggregating co-occurring lexical and phrasal micro-features into macro-features (i.e., dimensions). It then validates these macro-features by testing them on three language proficiency criteria: L2 writing proficiency, lexical proficiency scores, and longitudinal lexical growth.

METHOD

Corpora Selection

We used three corpora in this study: The Yonsei English Learner Corpus (YELC; Rhee & Jung, 2014), the Crossley written lexical proficiency corpus (Crossley et al., 2011a), and the Salsbury longitudinal spoken corpus (Salsbury et al., 2011). YELC was chosen to develop lexical sophistication components because of its large sample size and because it is comprised of L2 writing samples that have been scored in terms of essay quality. We selected the Crossley lexical proficiency corpus because it contained both L1 and L2 writing samples that had been scored specifically for lexical proficiency, which allowed us to test the developed components in terms of lexical sophistication. The Salsbury corpus was chosen because it contained language samples from six participants collected over the course of a year that allowed us to track development of lexical sophistication over time. Each corpus is briefly discussed below.

YELC is a Korean English as a foreign language (EFL) learner corpus. It consists of 6,572 texts with 1,085,828 words written by 3,286 Korean high school graduates (or equivalents). Each student's writings include one narrative essay and one argumentative essay. Student writings are evaluated by native speakers of English who receive training sessions. The writings are graded according to the Common European Framework of Reference for Languages (CEFR). The six levels of proficiency in the CEFR are subdivided into nine proficiency levels in YELC: A1 (beginner), A1+, A2, B1, B1+, B2, B2+, C1, and C2 (mastery). In the study, the L2 writing proficiency levels are transformed into scores from A1 = 1 to C2 = 9. The corpus used in the study comprises 3,286 argumentative essays at 770,511 words. Based on Crossley and McNamara's (2013) corpus selection criteria, 255 argumentative texts that fell beneath the 100-word threshold were excluded.[4] The remaining texts ($n = 3{,}031$) were used in this analysis. YELC was also used to test the validity of the developed component scores to predict L2 writing proficiency levels.

The Crossley written lexical proficiency corpus consists of 240 written samples from L1 and L2 English speakers. A total of 180 L2 written samples were collected from English language learners in an intensive language program at a university in the United States. They included 60 samples for three proficiency levels (i.e., beginning, intermediate, and advanced). The L2 samples were unstructured and naturalistic daily written journals that were completed as part of the participants' intensive English coursework. The L2 learners, whose age ranged from 18 to 27 years old, were from various L1 backgrounds (Korean, Japanese, Arabic, French, Bambara, Portuguese, Spanish, and Turkish). A comparison corpus of 60 free writings produced by native English-speaking college students was selected from the Stream of Consciousness Data Set from the Pennebaker Archive Project (Newman et al., 2008). All of the 240 written samples were assessed by three expert raters who were native speakers of English. The raters used a 5-point scaled, holistic rubric for

lexical proficiency (see Crossley et al., 2011a, for the rubric). The lexical proficiency scores ranged from 5 (i.e., skillful use of a varied, accurate, and apt vocabulary with ease and fluency) to 1 (i.e., limited vocabulary, incorrect word choices, and the use of words that obscure meaning comprehension). This corpus was used to validate the developed components by using them to predict human ratings of lexical proficiency.

The Salsbury longitudinal spoken corpus consists of 99 spoken samples collected over 1 year from six L2 English language learners in an intensive English program at a university in the United States. The six L2 learners, whose age ranged from 18 to 29 years old, comprised three Arabic speakers, one Spanish speaker, one Japanese speaker, and one Korean speaker. The learners were all placed in a beginning level class upon entry. They were interviewed every 2 weeks (excluding school breaks) over the course of 1 year. The interviewers used various elicitation materials, and the L2 learners and interviewers were allowed to simultaneously introduce their own conversation topics. The interviews lasted from 30 to 45 minutes and were tape-recorded and transcribed. We used this corpus to assess whether the developed components could track lexical growth over a year-long period.

Indices of Lexical and Phrasal Sophistication

A total of 424 indices of lexical sophistication taken from TAALES 2.0 were included as candidate indices in the study (see Kyle & Crossley, 2015). Each category is briefly discussed below.

Word Frequency. Frequency indices in TAALES are calculated based on various corpora including the Lorge's corpus of popular magazine articles (Thorndike & Lorge, 1944), the Brown Corpus (Kučera & Francis, 1967), the London–Lund Corpus of Conversation (Brown, 1984), two BNC sub-corpora (i.e., written and spoken; BNC, 2007), SUBTLEXus (Brysbaert & New, 2009), five Corpus of Contemporary American English (COCA) subcorpora (i.e., academic, fiction, magazine, news, and spoken; Davies, 2008), and the Hyperspace Analogue to Language (HAL) corpus (Balota et al., 2007). Each index calculates the mean word frequency scores by dividing the sum of the frequency scores for the words in a text by the number of words in that text that receive a frequency score. Logarithmic frequency scores are also calculated. If a word in the text is not in the selected frequency lists, it is not included in the mean frequency counts. Index variants include all words (AW), content words (CW), function words (FW), raw scores, and log-transformed scores.

Word Range. TAALES calculates range indices for the Brown Corpus, SUBTLEXus, two BNC subcorpora, five COCA subcorpora, and the Brown Corpus. TAALES calculates the mean word range scores by dividing the sum of the range scores for the words in a text by the number of words in that text that receive a range score. Index variants include AW, CW, FW, raw scores, and log-transformed scores.

Contextual Distinctiveness. Contextual distinctiveness norms in TAALES are based on various association values reported in the Edinburgh Associative Thesaurus (EAT; Kiss et al., 1973), the University of South Florida stimuli count index (Nelson, McEvoy, & Schreiber, 2004), McDonald's co-occurrence probability (McDonald & Shillcock, 2001), Hoffman's semantic diversity index (Hoffman, Ralph, & Rogers, 2013), and Latent Semantic Analysis values (Landauer et al., 2007).

Psycholinguistic Norms. Psycholinguistic norms for words in TAALES are derived from the MRC Psycholinguistic Database (Coltheart, 1981) familiarity, imageability, concreteness, and meaningfulness. Additional AoA indices are based on Kuperman et al. (2012), and additional concreteness indices are based on Brysbaert et al. (2014). High scores reflect texts containing more familiar, imageable, concrete, and meaningful words, while for AoA, high scores reflect texts containing words that are acquired at a later age.

Word Neighborhood. Word neighborhood norms in TAALES are taken from the English Lexicon Project (ELP; Balota et al., 2007), which contains word neighborhood information of 40,481 words. It includes word neighborhood size and frequency indices for orthographic, phonographic, and phonological neighbors. Neighborhood frequency norms are based on the HAL corpus (Lund & Burgess, 1996).

Word Recognition. Lexical decision and word naming behavioral norms in TAALES are taken from the ELP. The ELP includes native English speakers' response times and accuracies to 40,481 words during lexical decision and word naming tasks.

Age of Exposure. AoE indices in TAALES are based on incremental AoE for words across 13 grade levels taken from the TASA corpus (Dascalu et al., 2016). High AoE scores reflect texts containing words in higher grade levels, while lower

AoE scores reflect texts containing words in lower grade levels.

Semantic Relations. Hypernymy and polysemy norms in TAALES are based on WordNet (Fellbaum, 1998) in which over 170,000 English nouns, verbs, adjectives, and adverbs are organized according to connections between relevant lexical concepts. Hypernymy scores represent the number of superordinate terms a word has while polysemy scores represent the number of senses a word has. Hypernymy index variants include nouns, verbs, and both nouns and verbs. High hypernymy scores reflect texts containing less abstract and more specific words. Polysemy indices are reported for all words. High polysemy scores reflect texts containing words with more senses.

N-Gram Frequency, Range, and Proportion. Frequency, range, and proportion norms for bigrams and trigrams in TAALES are calculated based on BNC and COCA. Each frequency/range index calculates the mean *n*-gram frequency/range scores by dividing the sum of the frequency/range scores for the *n*-grams in a text by the number of *n*-grams in that text that receive frequency/range scores. The proportion of *n*-grams in the target text that also occur frequently in the reference corpus (e.g., are among the 50,000 most frequent *n*-grams) is also calculated.

N-Gram Association Strength. TAALES includes five association strength norms for bigrams and trigrams which are derived from COCA: MI score, MI Squared (MI^2) score, *t*-score, ΔP score, and approximate collexeme score.[5] Each *n*-gram association strength index is calculated by dividing the sum of the association strength scores for *n*-grams in a text by the number of *n*-grams in that text that receive the association strength score.

Academic Language. Academic language norms in TAALES are based on word-level academic lists (Coxhead, 2000) and phrase-level academic lists (Simpson–Vlach & Ellis, 2010). Academic language indices are calculated by counting the number of tokens in a text that also occur in an academic list and then dividing that number by the number of words in the text.

Statistical Analyses

We conducted four studies to answer each research question. In Study 1, we investigated whether lexical micro-features could be grouped into lexical macro-features (RQ1). To accomplish this, we conducted a principal component analysis (PCA), which is a statistical procedure that clusters variables that are highly correlated with each other into groups, such that a number of correlated variables can be reduced into a smaller set of derived variables (i.e., components or dimensions; Jolliffe, 2002). The PCA allowed us to reduce a large number of lexical sophistication indices selected from TAALES to a smaller set of components or dimensions. Prior to the PCA, we checked for normal distribution of the TAALES indices by measuring their skewness and kurtosis levels. Only those variables that were normally distributed were included in the PCA. We next controlled for multicollinearity between indices (defined as $r > .90$). If two of more indices were multicollinear, only one of the indices was included in the PCA. For the PCA, we calculated *z*-scores for each index to standardize scores across indices. For inclusion into a component, a conservative cut-off for the eigenvalues of $\lambda > .40$ was set, which ensured that salient lexical sophistication indices would be included in the components. If an index loaded into several components, the index was only included in the component in which it loaded highest. Based on the PCA results, we calculated weighted component scores by multiplying each lexical index by the eigenvalue of the same index reported by the PCA.

In Studies 2 and 3, we performed regression analyses to investigate whether the component scores developed in Study 1 could account for the variance of L2 writing proficiency levels as reported in the YELC (Study 2; RQ2) and holistic lexical proficiency scores as reported in the Crossley written lexical proficiency corpus (Study 3; RQ3). We first ran the texts in the corpora through TAALES and then calculated *z*-scores for each index included in the components derived in Study 1. We then calculated weighted component scores in YELC and the Crossley corpus based on the PCA from Study 1. We then used SPSS to approximately divide the corpus into a training set and a test set following a 67/33 split (Witten, Frank, & Hall, 2011). Components that showed significant correlations with L2 writing proficiency levels and holistic lexical proficiency scores in the training sets were included as predictor variables in a stepwise multiple regression to explain the variance of L2 writing proficiency levels and holistic lexical proficiency scores, respectively. The regression models that were derived from these analyses were then applied to the essays in the test set to assess how well the model worked on predicting an independent set.

In Study 4 we investigated whether lexical macro-features (i.e., the component scores) could be used to track L2 learners' longitudinal

development (RQ4). To accomplish this, we performed repeated measures analyses of variance (RM ANOVAs). We calculated *z*-scores for each index included in the components and weighted component scores in the Salsbury corpus. From the total of 99 spoken samples from six participants, we first chose the weeks when every participant's spoken samples were collected, and then selected the weeks in which component scores were normally distributed. This left us with six sets of data over 1 year (i.e., the 2nd, 11th, 24th, 32nd, 48th, and 52nd weeks after the participants' arrival in the United States). In the 32nd week, one component was not normally distributed and was excluded from analyses.

RESULTS

Study 1: Principal Component Analysis for Grouping Lexical Micro-Features Into Lexical Macro-Features

In our first study we developed component scores for the TAALES indices. Of the 424 lexical sophistication indices reported in TAALES, 56 indices were not normally distributed and were removed from the analysis. Of the remaining 368 indices, 268 indices showed strong multicollinearity and were removed, retaining 100 indices for the PCA. For the PCA, a Promax rotation method (i.e., correlated solution) with 25 maximum iterations for convergence was used. The Kaiser–Meyer–Olkin test indicated that the measuring of sampling adequacy was acceptable (0.89). Based on the PCA results, we included the first 12 components whose eigenvalues were above 1.5. The 12 components included a total of 90 lexical sophistication indices. The first 12 components accounted for 77.59% of the shared variance in YELC (see Table 1). Each component was assigned a name that best captured the characteristics of the indices that loaded into it. Table 1 presents each component along with scores of each index that loaded into it for example words and *n*-grams.

TABLE 1
Shared Variance of the First 12 Components From the Principal Component Analysis

Component (C) Number and Name	Percent of Variance	Cumulative Variance
C1: Bigram and trigram proportions	22.845	22.845
C2: Word acquisition properties	16.049	38.894
C3: Content word frequency	11.495	50.389
C4: Trigram mutual information	5.683	56.072
C5: Bigram frequency and range	4.405	60.477
C6: Function word properties	3.936	64.413
C7: Content word properties	3.126	67.539
C8: Bigram mutual information	2.380	69.919
C9: Function word frequency and range	2.175	72.094
C10: Bigram and trigram strength of directional association	2.039	74.132
C11: Academic formulaic language	1.833	75.966
C12: Word specificity	1.622	77.588

The first component (*Bigram and Trigram Proportions*) encompassed bigram and trigram proportion scores based on BNC and COCA (see Table 2). This component included higher proportions of bigrams and trigrams found in the reference corpora *n*-gram lists. It also included a word frequency index and a word range index, containing words that are more frequent in the reference corpus as a whole and words that are used in more texts across the reference corpus. Regarding example *n*-grams as shown in Table 2, it should be noted that, although word frequency values for "*is able to*" and "*is able for*" are similar, they differ in *n*-gram frequency: the former contains two frequent bigrams (i.e., *is able* and *able to*) that also comprise a frequent trigram, while the latter contains one frequent bigram (i.e., *is able*) and one infrequent bigram (i.e., *able for*) that together comprise an infrequent trigram. Thus, "*is able to*" has greater values in bigram and trigram proportion measures than "*is able for.*"

The second component (*Word Acquisition Properties*) comprised word acquisition properties (see Table 3). The indices that loaded positively into this component included word frequency and range, orthographic neighbors with higher frequency, phonological neighbors, and phonographic neighbors. The indices that loaded negatively into this component included orthographic neighbors with lower frequency, naming response time, lexical decision time, academic word lists, and age of exposure.

The indices that loaded into the third component *(Content Word Frequency)* included content

TABLE 2
Component 1: Bigram and Trigram Proportions

Index	Eigen Loading	Example 1: *is able to*	Example 2: *is able for*
BNC Written Bigram Proportion 50k	.937	1	.5
COCA Magazine Bigram Proportion 10k	.925	.5	0
COCA Fiction Bigram Proportion 60k	.896	1	.5
COCA News Bigram Proportion 90k	.880	1	.5
COCA Academic Bigram Proportion 10k	.878	1	.5
COCA Academic Bigram Proportion 100k	.858	1	.5
COCA Spoken Trigram Proportion 20k	.818	1	0
BNC Written Trigram Proportion 50k	.813	1	0
COCA Fiction Frequency AW Logarithm	.776	3.458	3.267
COCA Academic Trigram Proportion 30k	.769	1	0
COCA Fiction Trigram Proportion 30k	.764	1	0
COCA Magazine Trigram Proportion 10k	.754	1	0
COCA Spoken Range AW Logarithm	.741	−.130	−.131
COCA Fiction Trigram Proportion 10k	.650	0	0
Component score		11.018	4.663

TABLE 3
Component 2: Word Acquisition Properties

Index	Eigen Loading	Example 1: *people*	Example 2: *admission*
SUBTLEXus Range CW	.874	7889	164
Orthographic Neighbors With Higher Frequency (mean frequency)	.851	7.02	7.42
Phonological Neighbors (excludes homonyms)	.835	7	0
BNC Spoken Range AW	.834	92.509	4.530
SUBTLEXus Range AW Logarithm	.827	3.897	2.218
Orthographic Neighbors	.820	0	0
Thorndike–Lorge Frequency CW Logarithm	.782	3.553	1.845
Phonographic Neighbors (homophones excluded)	.727	0	0
Orthographic Neighbors With Lower Frequency (mean number)	−.873	1.8	2.05
Word Naming Response Time	−.798	573.6	637.04
LDA Age of Exposure (.40 cosine threshold)	−.633	0	5
Lexical Decision Time	−.593	540.85	694.71
Academic Word List All	−.552	0	0
LDA Age of Exposure (inverse slope)	−.507	.793	1.638
Component score		6209.529	−769.401

word frequency scores based on various reference corpora (see Table 4). The component loaded content words that are more frequent in each reference corpus as a whole. It also included a content word range index, encompassing content words that are used in more texts across the reference corpus.

The fourth component (*Trigram Mutual Information)* represented trigram MI indices based on COCA (see Table 5). The trigram MI indices calculated association strength between the first bigram and the remaining word in a trigram. The component included higher trigram MI scores, which indicates that trigram sequences tend to be made up of low-frequency words.

The fifth component (*Bigram Frequency and Range*) included bigrams that are more frequent in each reference corpus as a whole and bigrams

TABLE 4
Component 3: Content Word Frequency

Index	Eigen Loading	Example 1: *begin*	Example 2: *investigate*
COCA Spoken Frequency CW	.908	233.729	20.932
SUBTLEXus Frequency CW	.906	2906	483
Brown Frequency AW	.875	14	1
Thorndike–Lorge Frequency CW	.860	1109	110
COCA Magazine Range CW	.850	0.163	0.02
COCA Fiction Frequency CW	.839	90.7	8.813
BNC Spoken Frequency CW Logarithm	.830	−1.189	−2.198
Brown Frequency CW Logarithm	.817	1.146	0
BNC Written Frequency CW Logarithm	.801	−1.093	−1.591
COCA Fiction Frequency CW Logarithm	.794	1.958	0.945
COCA Academic Frequency CW	.776	109.983	38.272
Component score		3973.261	586.843

TABLE 5
Component 4: Trigram Mutual Information

Index	Eigen Loading	Example 1: *in order to*	Example 2: *in order for*
COCA Magazine Trigram Bigram to Unigram Association Strength (MI^2)	.902	12.246	6.977
COCA News Trigram Bigram to Unigram Association Strength (MI^2)	.894	11.630	6.790
COCA Fiction Trigram Bigram to Unigram Association Strength (MI^2)	.844	11.493	6.055
COCA Spoken Trigram Bigram to Unigram Association Strength (MI^2)	.828	12.303	8.278
COCA Magazine Trigram Bigram to Unigram Association Strength (MI)	.827	3.504	1.374
COCA Spoken Trigram Bigram to Unigram Association Strength (MI)	.784	3.454	2.070
COCA Fiction Trigram Bigram to Unigram Association Strength (MI)	.784	3.462	1.401
COCA Academic Trigram Bigram to Unigram Association Strength (MI^2)	.767	13.206	7.879
COCA Magazine Trigram Bigram to Unigram Association Strength (MI^2)	.902	12.246	6.977
COCA News Trigram Bigram to Unigram Association Strength (MI^2)	.894	11.630	6.790
COCA Fiction Trigram Bigram to Unigram Association Strength (MI^2)	.844	11.493	6.055
Component score		59.781	34.228

that are used in more texts across each reference corpus (see Table 6). Additionally, it included a word frequency index and two bigram MI indices based on COCA, containing words that are more frequent in the reference corpus, and bigrams that are higher MI scores.

The indices that loaded positively into the sixth component (*Function Word Properties)* included concreteness and meaningfulness for function words, and concreteness of content words (see Table 7).[6] The indices that loaded negatively into the component included function word range and frequency.

The indices that loaded positively into the seventh component (*Content Word Properties)* included imageability and concreteness for content words, and imageability, concreteness, and meaningfulness for all words (see Table 8). The index that loaded negatively included age of acquisition for content words.

The eighth component (*Bigram Mutual Information*) captured bigram mutual information (see

TABLE 6
Component 5: Bigram Frequency and Range

Index	Eigen Loading	Example 1: *look at*	Example 2: *account for*
COCA Magazine Bigram Range	.901	.154	.03
COCA Fiction Bigram Range	.866	.399	.018
BNC Written Bigram Frequency Logarithm	.853	−.955	−1.658
COCA Academic Bigram Range	.851	.149	.109
COCA Academic Bigram Frequency Logarithm	.793	1.766	1.629
COCA Magazine Bigram Frequency Logarithm	.761	2.084	1.32
BNC Written Bigram Frequency	.756	.111	.022
HAL Frequency	.672	1115705.5	2509497
COCA Fiction Bigram Association Strength (MI^2)	.655	13.544	8.766
COCA Magazine Bigram Association Strength (MI^2)	.620	13.11	10.784
Component score		749773.962	1686395.446

TABLE 7
Component 6: Function Word Properties

Index	Eigen Loading	Example 1: *her*	Example 2: *of*
Brysbaert Concreteness Combined FW	.866	3	1.670
MRC Concreteness FW	.851	419	180
MRC Meaningfulness FW	.837	507	217
COCA Academic Range FW	−.747	.449	1
BNC Written Frequency FW	−.637	3.391	32.963
Component score		781.030	314.511

Note. FW = function words.

TABLE 8
Component 7: Content Word Properties

Index	Eigen Loading	Example 1: *car*	Example 2: *equality*
MRC Imageability AW	.884	638	346
MRC Imageability CW	.870	638	346
Brysbaert Concreteness Combined CW	.831	4.89	1.41
Brysbaert Concreteness Combined AW	.768	4.89	1.41
MRC Meaningfulness AW	.620	553	443
Age of Acquisition CW	−.524	3.37	10.27
Component score		1467.965	878.417

Note. AW = all words; CW = content words.

Table 9). The component included higher bigram MI scores, which indicated that bigrams tended to be made up of low-frequency words.

The ninth component (*Function Word Frequency and Range*) loaded function word range and frequency indices, containing function words that are more frequent and function words that are used in more contexts (see Table 10).

The tenth component (*Bigram and Trigram Strength of Directional Association*) included five ΔP indices based on COCA: three trigram ΔP indices (i.e., the probability of the third word in a trigram given the occurrence of the first bigram in it) and two bigram ΔP indices (i.e., the probability of the second word in a bigram given the occurrence of the first word in it; see Table 11).

TABLE 9
Component 8: Bigram Mutual Information (MI)

Index	Eigen Loading	Example 1: *last year*	Example 2: *nuclear weapon*
COCA Fiction Bigram Association Strength (MI)	.886	4.966	7.687
COCA Academic Bigram Association Strength (MI)	.867	5.265	7.709
COCA Spoken Bigram Association Strength (MI)	.827	5.145	7.393
Component score		13.220	19.608

TABLE 10
Component 9: Function Word Frequency and Range

Index	Eigen Loading	Example 1: *her*	Example 2: *of*
COCA Magazine Range FW	.855	.448	.998
COCA Fiction Frequency FW Logarithm	.851	4.004	4.294
BNC Spoken Range FW	.839	72.997	99.652
SUBTLEXus Range FW	.791	8076	8375
COCA Academic Frequency FW Logarithm	.752	2.987	4.584
Component score		6455.397	6716.188

Note. FW = function word.

TABLE 11
Component 10: Bigram and Trigram Strength of Directional Association

Index	Eigen Loading	Example 1: *in order to*	Example 2: *in order for*
COCA Spoken Trigram Bigram to Unigram Association Strength (ΔP)	.833	.822	.053
COCA Fiction Trigram Bigram to Unigram Association Strength (ΔP)	.806	.758	.020
COCA Academic Trigram Bigram to Unigram Association Strength (ΔP)	.790	.875	.029
COCA Fiction Bigram Association Strength (ΔP)	.696	.141	.011
COCA Academic Bigram Association Strength (ΔP)	.677	.221	.012
Component score		2.235	.099

The eleventh component (*Academic Formulaic Language*) included two trigram indices based on the COCA academic sub-corpus, and two academic formulaic sequence indices based on the AFL (see Table 12).

The indices that loaded positively into the twelfth component (*Word Specificity*) included word hypernymic scores for verbs, such that the component comprised verbs with higher hypernymic scores (i.e., use of more specific verbs; see Table 13). The indices that loaded negatively into this component included content word frequency scores and orthographic neighborhood frequency scores, such that the component included content words that are used less frequently, and words that have less orthographic neighbors.

Study 2: Predicting L2 Writing Proficiency Using Lexical Sophistication Components

In the second study, we tested the developed component scores' ability to explain L2 writing proficiency levels in the YELC. For the training set (n = 2,024), 11 components reported significant correlations with L2 writing proficiency

TABLE 12
Component 11: Academic Formulaic Language

Index	Eigen Loading	Example 1: *in the context of*	Example 2: *in order for*
COCA Academic Trigram Range Logarithm	.844	−.906	−1.544
COCA Academic Trigram Unigram to Bigram Association Strength (MI^2)	.792	11.165	9.994
Academic Formulas List All	.708	1	0
Academic Formulas List Core	.637	1	0
Component score		9.423	6.612

Note. MI = mutual information.

TABLE 13
Component 12: Word Specificity

Index	Eigen Loading	Example 1: *begin*	Example 2: *investigate*
Hypernymy Verbs (Sense Mean, Path Mean)	.791	.9	2
Hypernymy Verbs (Sense 1, Path 1)	.778	1	2
COCA Magazine Frequency CW Logarithm	−.804	2.113	1.138
Kucera–Francis Frequency CW Logarithm	−.739	1.924	1.041
Orthographic Neighborhood Frequency	−.469	6.89	0
Component score		−2.941	.090

Note. CW = content word.

levels (see Table 14). A regression analysis using the 11 significant weighted component scores as predictors for L2 writing proficiency levels yielded a significant model, $F(7,2016) = 93.782$, $p < .001$, $r = .496$, $R^2 = .246$. The regression model demonstrated that the seven component scores together explained 24.6% of the variance in 2,024 students' L2 writing proficiency levels in the training set (see Table 15 for details). The results indicated that higher proficiency writing was characterized by the production of bigrams and trigrams that on average are more strongly associated, a greater proportion of common bigrams and trigrams, and a greater number of advanced content words (e.g., words that occur less frequently; words that are used in fewer contexts; words that are less imageable, less concrete, and less meaningful; words that are less phonologically and phonographically dense; words that elicit longer response times; verbs that are more specific; and words that are acquired at a later age).

The regression model reported for the training set was used to predict L2 writing proficiency levels in the test set ($n = 1{,}007$). The regression model applied to the test set reported $r = .497$, $R^2 = .247$. Thus, the seven weighted components explained 24.7% of the variance in 1,007 students' L2 writing proficiency levels in the test set. These results indicate that the seven-component model identified in the initial stepwise multiple regression is stable across the data and generalizable to other student populations.

Study 3: Predicting Lexical Proficiency Using Lexical Sophistication Components

For the third study, we used the components from Study 1 to predict human ratings of lexical proficiency as reported in the Crossley written lexical proficiency corpus. In the training set ($n = 169$), eight components reported significant correlations with holistic lexical proficiency scores (see Table 14). A regression analysis using the eight significant weighted component scores as predictors for holistic lexical proficiency scores yielded a significant model, $F(5,163) = 14.625$, $p < .001$, $r = .556$, $R^2 = .310$. The regression model demonstrated that the five component scores together explained 31.0% of the variance in holistic lexical proficiency scores in the training set (see Table 16 for details). The results indicate that writers judged to be more lexically proficient used more strongly associated bigrams and trigrams, a greater proportion of common bigrams and trigrams, and a greater number of advanced content words (e.g., words that occur less frequently;

TABLE 14
Correlations Between Component Scores and L2 Writing Proficiency and Between Component Scores and Lexical Proficiency in the Training Sets for Regression Analyses

Component (C)	Correlation With L2 Writing Proficiency	Correlation With Lexical Proficiency
C1: Bigram and trigram proportions	.190**	.241**
C2: Word acquisition properties	−.226**	−.149
C3: Content word frequency	−.239**	−.072
C4: Trigram mutual information	−.159**	.320**
C5: Bigram frequency and range	.210**	.149
C6: Function word properties	−.246**	−.157*
C7: Content word properties	−.213**	−.306**
C8: Bigram mutual information	.204**	.214**
C9: Function word frequency and range	.059	.065
C10: Bigram and trigram strength of directional association	.295**	.384**
C11: Academic formulaic language	.186**	.164*
C12: Word specificity	.230**	.168**

Note. ** indicates $p < .01$; * indicates $p < .05$.

words that are less imageable; less concrete, and less meaningful; words that are acquired at a later age; and verbs that are more specific).

The regression model applied to the test set ($n = 71$) reported $r = .558$, $R^2 = .311$. Thus, the five weighted components explained 31.1% of the variance in holistic lexical proficiency scores in the test set, providing increased confidence for the generalizability of our five-component model.

Study 4: Tracking Longitudinal Lexical Growth Using Lexical Sophistication Components

To explore whether beginning L2 learners' lexical production develops over the course of 1 year in terms of the derived component scores, we conducted RM ANOVAs using the 12 components (see Table 17). Results indicated significant positive linear trends between time and four components with large effects: *bigram and trigram proportions* ($p < .001$, $\eta^2_p = .886$), *word acquisition properties* ($p < .001$, $\eta^2_p = .889$), *content word frequency* ($p < .05$, $\eta^2_p = .723$), and *bigram frequency and range* ($p < .05$, $\eta^2_p = .695$). The results also indicated significant negative linear trends between time and two components with large effects: *content word properties* ($p < .05$, $\eta^2_p = .733$), and *function word frequency and range* ($p < .001$, $\eta^2_p = .827$). The results indicate that over time, L2 learners began to produce a greater proportion of common *n*-grams; bigrams that are more frequently used; content words that are more frequently used; words that are less imageable, less concrete, and less meaningful; words that are acquired at a later age; and function words that are less frequently used.

DISCUSSION

Previous studies have focused on individual features of lexical sophistication and their relations to lexical development in L2 learners. The purpose of this study was to examine whether individual lexical features can be aggregated into components of lexical sophistication and to assess whether these components are predictive of lexical development in a number of tasks and domains. Such an approach can provide information about the multidimensional nature of lexical sophistication that may reflect learners' vocabulary knowledge. In this study, using the PCA, we reduced 100 micro-features of lexical sophistication taken from TAALES 2.0 into 12 dimensions. The 12 components include six word-level and six phrase-level components.

The PCA findings indicate strong overlap and connectivity among micro-features of lexical sophistication. For example, in the second component (*word acquisition properties*), various lexical micro-features that are conceptually and operationally distinct closely interacted with each other to form a larger dimension of lexical sophistication. These features included word range; word frequency; orthographic, phonological, and phonographic neighbors; response times to words; age of exposure; and academic words, which are all related to lexical acquisition but unrelated in terms of lexical properties. In another example, hypernymic relations were strongly related to word frequency indices in the 12th component (*word specificity*). The findings indicate that more specific verbs tend to also be less frequent. The overlap reported in many of

TABLE 15
Stepwise Regression Analysis for Component (C) Scores Predicting L2 Writing Proficiency

Entry	Components Included	*r*	R^2	R^2 change	β	*SE*	*B*
1	C10: Bigram and trigram strength of directional association	.295	.087	.087	.021	.009	.058
2	C1: Bigram and trigram proportions	.340	.116	.029	.044	.003	.360
3	C2: Word acquisition properties	.442	.196	.080	−.030	.004	−.212
4	C12: Word specificity	.475	.226	.030	.085	.012	.201
5	C7: Content word properties	.486	.236	.010	−.041	.008	−.123
6	C8: Bigram mutual information	.491	.241	.005	.038	.011	.077
7	C3: Content word frequency	.496	.246	.004	−.015	.004	−.102

Note. Estimated constant term = 4.326; β = unstandardized beta; *SE* = standard error; *B* = standardized beta.

TABLE 16
Stepwise Regression Analysis for Component (C) Scores Predicting Lexical Proficiency

Entry	Components Included	*r*	R^2	R^2 change	β	*SE*	*B*
1	C10: Bigram and trigram strength of directional association	.384	.148	.148	.068	.024	.209
2	C7: Content word properties	.441	.194	.047	−.046	.020	−.172
3	C8: Bigram mutual information	.471	.221	.027	.102	.031	.233
4	C1: Bigram and trigram proportions	.497	.247	.026	.035	.009	.324
5	C12: Word specificity	.556	.310	.062	.129	.034	.291

Note. Estimated constant term = 3.200; β = unstandardized beta; *SE* = standard error; *B* = standardized beta.

the derived components provides evidence for the interconnected nature of L2 lexical features (Crossley et al., 2011a, 2011b; Zareva, Schwanenflugel, & Nikolova, 2005).

These components were also predictive of a number of tasks related to lexical proficiency. For instance, a number of these component scores were predictive of human judgments of L2 writing proficiency levels. L2 writing performance involves learners' control over language use, including vocabulary, grammar, spelling, cohesive devices, and punctuation (Leki, Cumming, & Silva, 2008) and research has found that more proficient writers tend to use more sophisticated words, such as low-frequency words, words that occur in fewer contexts, less imageable words, less meaningful words, less familiar words, and words that are acquired at a later age (Crossley & McNamara, 2012; Crossley et al., 2014; Jung, Crossley, & McNamara, 2015; Kyle & Crossley, 2016; Laufer & Nation, 1995). Beyond word-level features, research has also shown that *n*-gram frequency is predictive of L2 writing scores such that L2 texts with more frequent *n*-grams likely receive higher scores (Jung et al., 2015; Kyle & Crossley, 2016). We found similar results in the current study in that writing quality was predicted by components related to sophisticated words (e.g., low-frequency words, less familiar words, words that are acquired at a later age) and *n*-grams (e.g., *n*-grams that are frequently used).

Likewise, the derived components were predictive of human judgments of lexical proficiency. Previous studies have demonstrated that L2 learners who produce more sophisticated words and *n*-grams are judged to be more lexically proficient (Crossley et al., 2009, 2010, 2011a, 2011b, 2013; Crossley, Kyle, & McNamara, 2015; Kyle & Crossley, 2015). Similarly, the component analysis in this study also reports that more proficient L2 learners produced language that was more sophisticated in that it included a greater proportion of frequent *n*-grams and a greater number of advanced content words (e.g., low-frequency words, words that are less concrete, words that are acquired at a later age, and verbs that are more specific).

Of interest is the overlap between the L2 writing quality regression model and the lexical proficiency model in that the two regression models shared five components as predictors (i.e., *bigram and trigram strength of directional association, content word properties, bigram mutual information,*

TABLE 17
Mean (Standard Deviation) for Component (C) Scores Across the Year

Week	2	11	24	32	48	52	F	η^2_p
C1: Bigram and trigram proportions	−14.724 (11.021)	−5.385 (9.461)	−5.169 (9.074)	.570 (10.842)	6.674 (7.008)	7.438 (5.197)	38.709**	.886
C2: Word acquisition properties	−11.975 (8.353)	−3.216 (7.726)	−4.701 (8.842)	3.864 (9.032)	3.552 (6.234)	4.374 (6.307)	39.933**	.889
C3: Content word frequency	−8.148 (6.132)	−3.623 (5.604)	−2.920 (9.227)	.073 (3.637)	.469 (6.008)	4.796 (5.132)	13.065*	.723
C4: Trigram mutual information	−3.007 (4.458)	−3.005 (7.550)	4.624 (8.003)	−.019 (3.784)	2.759 (3.351)	3.059 (3.506)	5.613	n/a
C5: Bigram frequency and range	−7.675 (6.740)	−1.336 (7.279)	−1.549 (3.301)	.090 (6.919)	3.301 (4.537)	2.592 (4.650)	11.370*	.695
C6: Function word properties	.173 (3.742)	−.074 (2.727)	.068 (2.843)	2.172 (2.530)	.626 (3.798)	1.220 (4.059)	.696	n/a
C7: Content word properties	3.750 (2.929)	2.222 (2.251)	1.197 (2.264)	−1.094 (2.015)	−2.416 (2.940)	−2.786 (3.712)	13.698*	.733
C8: Bigram mutual information	−.863 (4.361)	−1.475 (1.217)	.654 (2.665)	.716 (2.140)	1.239 (1.639)	1.301 (1.652)	3.411	n/a
C9: Function word frequency and range	4.400 (3.389)	1.173 (2.512)	−.231 (2.441)	.071 (1.931)	−.782 (2.957)	−2.269 (2.301)	23.923**	.827
C10: Bigram and trigram strength of directional association	1.146 (3.447)	−1.633 (3.659)	1.619 (4.190)	non-normal	1.177 (2.497)	1.161 (2.364)	.444	n/a
C11: Academic formulaic language	−1.274 (1.660)	−.185 (3.244)	−.345 (1.519)	−.186 (1.602)	.487 (0.965)	.381 (2.017)	6.331	n/a
C12: Word specificity	1.434 (3.413)	−.385 (.987)	0.691 (2.413)	−.237 (.816)	1.558 (1.349)	−.472 (1.513)	1.833	n/a

Note. ** indicates $p < .01$; * indicates $p < .05$.

bigram and trigram proportions, and *word specificity*). Such convergence is of interest because the two regression models concerned two different language abilities (i.e., L2 writing proficiency and lexical proficiency) using two different types of corpora (i.e., a corpus of EFL learners' argumentative writings and a corpus of free writings by L1 speakers and learners of English as a second language [ESL]), but included similar components as predictors of performance. This finding provides some evidence that some components may contain core lexical sophistication features that are key for L2 lexical research. In addition, the strongest predictor in both regression models was *bigram and trigram strength of directional association,* such that L2 learners who produce bigrams and trigrams whose directional associations are stronger will be judged to be more proficient. These results support the importance of multi-word units as a whole in language learning processes (Gries, 2013; Hoey, 2005; Sinclair, 1991). *Word specificity* was also a strong predictor in each regression model indicating that more proficient L2 learners are likely to produce more specific verbs. These results contrast with some previous studies (Crossley et al., 2009, 2011a), while supporting others (Ijaz, 1986). The regression models also indicated that more proficient L2 learners tend to use a higher proportion of *n*-grams found in the reference corpora *n*-gram lists (Kyle & Crossley, 2015), that the more strongly associated bigrams a L2 learner uses, the more likely the learner is to be judged as being more proficient, and that more proficient L2 learners tend to use content words that are less imageable and less concrete, and content words that are acquired at a later age (Crossley et al., 2014; Saito et al., 2016; Salsbury et al., 2011).

Overall, the two regression analyses demonstrated that the macro-features of lexical sophistication together explained 24.6% and 31.0% of the variance in L2 writing proficiency levels and holistic lexical proficiency scores, respectively. Some of the unexplained variance in L2 writing proficiency levels may be related to other linguistic elements, such as syntactic complexity, and textual elements, such as coherence, rhetoric, and logical development, all of which are potential predictors of writing quality (e.g., Leki et al., 2008). With reference to lexical proficiency,

the remaining variance in human judgments of lexical proficiency may be related to other elements important for predicting lexical proficiency such as fluency, comprehensibility, lexical diversity, proportion of lexical errors, and L2 learners' metacognitive strategies for lexical use (Chapelle, 1994; Iwashita et al., 2008; Lu, 2012; Read, 2000; Saito et al., 2016). Furthermore, it should be noted that the variance in lexical proficiency explained by macro-features of lexical sophistication (i.e., 31.0%) is lower than the variance explained by micro-features of lexical sophistication reported in previous studies (Crossley et al., 2011a, 2011b; Kyle & Crossley, 2015). For instance, Kyle and Crossley (2015) found that five micro-features of lexical sophistication explained 47.5% of the variance in holistic lexical proficiency scores. Thus, while macro-features are useful in exploring productive vocabulary at a larger dimensional level, micro-features are more predictive of human judgments of lexical proficiency.

In reference to longitudinal development, the study demonstrates that beginning L2 learners' lexical development over a year-long period involves growth in lexical sophistication components (i.e., content word frequency, content word properties such as concreteness and imageability, function word frequency and range, and word acquisition properties) and phrase-level sophistication components (i.e., *n*-gram frequency and *n*-gram proportions). In addition, no evidence was found that L2 beginning learners become sensitive to *n*-gram association strength (Li & Schmitt, 2009), function word properties, academic formulaic language, or word specificity over a year-long L2 immersion experience.

Some similarities between the two regression analyses and the longitudinal analysis were found. Two components were predictors of L2 writing proficiency levels and holistic lexical proficiency scores and also showed significant longitudinal growth: *bigram and trigram proportions* and *content word properties*. While the former showed a positive growth trend and demonstrated that learners begin to use a higher proportion of *n*-grams that are frequently used, the latter showed a negative growth trend, indicating that as L2 learners become more proficient, they tend to produce content words that are less imageable, less concrete, and acquired at a later age (Kyle & Crossley, 2015). These two components showed the strongest relationships with indicators of lexical sophistication across the three different corpora used in these studies. This finding is of interest given that the three corpora used in this study differed in learning contexts (i.e., EFL and ESL learners), L1 backgrounds, task types (i.e., argumentative essays, free writings, and natural conversation), and modality (i.e., spoken and written language). It is likely, then, that these two lexical properties may reflect general growth patterns in lexical sophistication that can be applied to a wide range of L2 learners and task types.

The differences between the regression analyses and the longitudinal analysis merit some discussion as well. Specifically, two components (i.e., *word acquisition properties* and *content word frequency*) were negative predictors of L2 writing proficiency levels in the regression model but showed positive growth trends in the longitudinal analysis. These differences might be attributed to the dependent variable of interest (i.e., time versus human judgments of writing level). Previous studies have shown that beginning L2 learners begin to use words that are more frequent as a function of time spent learning English in L2 immersion contexts over a year (Crossley et al., 2010). However, human raters are more interested in links between lexical sophistication and writing quality regardless of time spent learning a language. Thus, human raters will judge texts as being of higher quality based on the use of less frequent words, words that are used in fewer contexts, words that have fewer phonological and phonographic neighbors, words that elicit more response times, and words that are exposed at a later age (Crossley et al., 2011a, 2011b; Kyle & Crossley, 2015).

CONCLUSION

The current study developed 12 dimensions of lexical sophistication by aggregating co-occurring micro-feature in a large-scale learner corpus. It further tested whether these macro-features of lexical sophistication were predictive of language abilities (i.e., L2 writing proficiency and L2 lexical proficiency) as well as indicative of longitudinal lexical growth. Our main finding is that lexical sophistication can be conceptualized as a multidimensional phenomenon at the macro level by both word and phrase components and that these components provide evidence for a detailed summarization of lexical sophistication. This summarization is as follows: The construct of lexical sophistication is informed by the production of (a) advanced content words, such as words that are less frequent, less concrete, less imageable, less orthographically and phonologically dense, more specific, exposed and acquired at a later age, and processed more slowly and (b) advanced bigrams

and trigrams, such as bigrams and trigrams that occur more frequently and whose directional associations are stronger.

The findings from this study also have important theoretical implications. For instance, lexical sophistication is not based on frequency information alone, but is also based on various pieces of lexical information such as word neighborhood density, familiarity, and *n*-gram frequency and association strength. These findings generally support cognitive approaches, which argue that linguistic units are based on form–meaning pairings of words and multi-units (Ellis & Larsen−Freeman, 2009; Goldberg, 2006; Langacker, 2007). Furthermore, the phrasal sophistication findings may be indicative of general cognitive learning processes that rely on establishing associations among frequently co-occurring words (Gries, 2013). From phraseological perspectives, the study provides evidence that multi-unit lexical items are important components in L2 writing proficiency, L2 lexical proficiency, and longitudinal lexical growth (Ellis, 2002; Hoey, 2005; Sinclair, 1991). In line with usage-based approaches, we also find evidence that frequency information is important at the phrase level as well as the word level (Gries & Ellis, 2015).

The reported findings require additional studies to provide supporting evidence of their usefulness in both theory and application. Additionally, there are some limitations to the approaches used in these studies. First, with respect to L2 writing proficiency, correlation and regression analyses were conducted with L2 writing proficiency levels, not with test scores. Future studies would benefit from using test scores in exploring relations between lexical features and writing proficiency. Second, while macro-features encompass large-scale aggregations of lexical sophistication, they may not be merited in a number of research and pedagogical instances where the fine-grained aspects of lexical and phrasal sophistication captured in micro-features may be warranted. Third, while L1-based word norms (e.g., frequency and lexical decision latencies) are predictive of L2 lexical performance, word norms derived from L2 learner corpora and behavior data merit further consideration. Fourth, our selected corpora did not always take into account differences in learners' L1 backgrounds, modality (i.e., written and spoken), and task types. Last, while using computational indices for measuring lexical sophistication can provide us with critical information on lexical proficiency and development, automatically evaluating the accuracy of learner production, including assessing the appropriate use of collocations, is not yet possible. More advanced computational analyses that can consider linguistic accuracy would merit ample consideration.

NOTES

[1] Word meaningfulness differs from word frequency, such that frequent words are not always words with high meaningfulness scores, and vice versa. For instance, articles (e.g., *a* and *the*) and prepositions (e.g., *to* and *in*) are high-frequency words, but low meaningfulness words (i.e., fewer word associations). In contrast, some words with high meaningfulness scores (e.g., *alcohol* and *punish*) are not high-frequency words.

[2] Word concreteness and imageability norms have been reported as highly correlated (Connell & Lynott, 2012). However, concreteness and imageability differ in that many concrete words are also imageable, while abstract words vary in imageability ratings (Connell & Lynott, 2012). In general, one caveat in using psycholinguistic norms of words, including concreteness and imageability norms, is that these norms do not consider the effects of L1–L2 cognates.

[3] In using AoA ratings, one caveat is that L1 speakers' ratings of AoA may not reflect L2 learners' exposure to the L2. However, AoA ratings have their advantages over frequency information (which is mainly based on materials for adult readers) in reflecting the cumulative frequency (i.e., the degree to which people have encountered words through exposure from childhood; Kuperman et al., 2012).

[4] Among the 255 argumentative texts (less than 100 words) that were excluded, 39 texts were A1, 95 were A1+, 86 were A2, 33 were B1, and two were B2. According to Crossley & McNamara (2013), setting the 100-word cutoff helps report more reliable values because a sample with a minimum of 100 words can provide enough lexical coverage to compute various lexical indices (e.g., frequency and concreteness). It should be noted that after operating the 100-word cutoff, the trimmed corpus still contained a large number of beginning-level texts (i.e., 691 texts from A1 to A2 levels).

[5] Mutual Information (MI) is calculated as the logarithm of the observed co-occurrence of two items divided by the expected co-occurrence of two items (Evert, 2005): MI = log(observed/expected). Mutual information squared (MI^2) scores that mitigate the emphasis of low-frequency items are calculated as the logarithm of the observed co-occurrence of two items (squared) divided by the expected co-occurrence of two items (Evert, 2005): $MI^2 = \log(\text{observed}^2/\text{expected})$. *T* is calculated as the observed frequency minus the expected frequency, divided by the square root of the observed frequency: $T = \frac{observed - expected}{\sqrt{observed}}$. Delta P is the probability of an outcome (O) given a cue (C) minus the probability of an outcome without the cue: delta P = P(O|C) – P(O|-C). Approximate collexeme strength that does not include normal distribution as an assump-

tion is calculated by multiplying the delta P value by the frequency of the first item.

[6] Some function words (e.g., *she*, *he*, and *up*) are rated as more concrete and meaningful than other function words (e.g., *a*, *the*, and *of*).

REFERENCES

Adelman, J. S., & Brown, G. D. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*, 455–459.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*, 147–149.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101–122). Oxford: Oxford University Press.

Balota, D. A, Yap, M. J., Cortese, M. J., Hutchison, K. A, Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459.

Berger, C. M., Crossley, S., & Kyle, K. (2017). Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second language production. *Applied Linguistics*. Early view: 13 March 2017, https://doi.org/10.1093/applin/amx005

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41.

BNC. (2007). *British National Corpus, version 3* (BNC XML ed.). Accessed 10 June 2017 at 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.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Chapelle, C. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*, 157–187.

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

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*, 452–465.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

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*, 250–270.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment*, *8*, 1–14.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*, 115–135.

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

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring second language lexical growth using hypernymic relationships. *Language Learning*, *59*, 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570–590.

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.

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.

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.

Daller, H., Milton, J., & Treffers–Daller, J. (Eds.). (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers–Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 150–164). Cambridge: Cambridge University Press.

Dascalu, M., McNamara, D. S., Crossley, S. A., & Trausan–Matu, S. (2016). Age of exposure: A model of word learning. In *The 30th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence* (pp. 2928–2934). Phoenix, AZ: AAAI Press.

Davies, M. (2008). *The Corpus of Contemporary American English: 520 million words, 1990–present.* Accessed 10 June 2016 at http://corpus.byu.edu/coca/

Douglas, R. D. (2013). The lexical breadth of undergraduate novice level writing competency. *Canadian Journal of Applied Linguistics, 16*, 152–170.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47*, 157–177.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*, 143–188.

Ellis, N. C., & Larsen–Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning, 59*(S1), 90–125.

Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations.* Unpublished doctoral dissertation. University of Stuttgart, Stuttgart, Germany.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1212–1248). Berlin/New York: Mouton de Gruyter.

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, MA: The MIT Press.

Forster, K. I., & Shen, D. (1996). No enemies in the neighborhood: Absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 696–713.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next … *International Journal of Corpus Linguistics, 18*, 137–166.

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning, 65*(S1), 228–255.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*, 218–238.

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* London: Routledge.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*, 718–730.

Ijaz, I. H. (1986). Linguistic and cognitive determinants of lexical acquisition in a second language. *Language Learning, 36*, 401–451.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 29–49.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(S1), 87–106.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Jung, Y., Crossley, S. A., & McNamara, D. S. (2015). *Linguistic features in MELAB writing performances* (Working Paper No. 2015-05). Cambridge Michigan Language Assessments. Accessed 18 September 2017 at http://www.cambridgemichigan.org/wp-content/uploads/2015/04/CWP-2015-05.pdf

Kaushanskaya, M., & Rechtzigel, K. (2012). Concreteness effects in bilingual and monolingual word learning. *Psychonomic Bulletin & Review, 19*, 935–941.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey, & N. Hamilton–Smith (Eds.), *The computer and literary studies* (pp. 153–165). Edinburgh, UK: Edinburgh University Press.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English. 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.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*, 757–786.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis.* Mahwah, NJ: Lawrence Erlbaum.

Langacker, R. W. (Eds.). (2007). Cognitive grammar. In D. Geeraets & H. Cuyckens, *The Oxford handbook of cognitive linguistics* (pp. 421–462). Oxford: Oxford University Press.

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

Leki, I., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English.* London: Routledge.

Levenston, E., & Blum, S. (1977). Aspects of lexical simplification in the speech and writing of advanced adult learners. In P. S. Corder & E. Roulet (Eds.), *The notions of simplification, inter-languages and pidgins and their relation to second language pedagogy* (pp. 51–72). Geneva, Switzerland: Librairie Droz.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing, 18*, 85–102.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal*, *96*, 190–208.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–322.

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: Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender difference in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*, 211–236.

Paivio, A. U. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Rhee, S.-C., & Jung, C. K. (2014). Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *Korea Contents Association*, *14*, 1019–1029.

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical profiles of comprehensible second language speech. *Studies in Second Language Acquisition*, *38*, 677–701.

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

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

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19*, 17–36.

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

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

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

Witten, I. H., Frank, E., & Hall, M. A., (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). San Francisco, CA: Morgan Kaufmann.

van Heuven, W., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effect in bilingual word recognition. *Journal of Memory and Language*, *39*, 458–483.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency—variable sensitivity. *Studies in Second Language Acquisition*, *27*, 567–595.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception & Performance*, *24*, 1131–1161.