Applied Linguistics 2017: 1–22
doi:10.1093/applin/amx005

Using Native-Speaker Psycholinguistic Norms to Predict Lexical Proficiency and Development in Second-Language Production

[1,*]CYNTHIA M. BERGER, [1]SCOTT A. CROSSLEY, and [2]KRISTOPHER KYLE

[1]Department of Applied Linguistics & ESL, Georgia State University and [2]Department of Second Language Studies, University of Hawai'i at Manoa
[*]E-mail: cberger@gsu.edu

A large data set of L1 psycholinguistic norms (Balota *et al.* 2007) was used to assess spoken L2 English lexical proficiency in cross-sectional and longitudinal learner corpora. Behavioral norms included lexical decision and word naming latencies (i.e. reaction times) and accuracies for 40,481 English words. A frequency measure was included to compare the relative strength of the norms to a traditional lexical measure when explaining proficiency and growth. The cross-sectional study revealed that learners identified as more lexically proficient by human raters produced words that were recognized more slowly and named more slowly and less accurately in L1 experimental settings. Moreover, lexical decision latencies explained more variance in ratings than frequency, while frequency and word naming latencies were comparable. The longitudinal study indicated that words produced by L2 speakers over time were recognized less accurately and named more slowly and less accurately by L1 subjects, while the frequency of those words decreased over time. Together, results demonstrate that L1 psycholinguistic information can index L2 lexical proficiency and growth across corpora while adding unique information to our understanding of L2 lexical knowledge.

Lexical proficiency remains an elusive construct that defies easy definition or quantification (Zareva *et al.* 2005; Boers *et al.* 2006; David 2008), despite its importance in second language (L2) learning (Alderson 2005; David 2008), reading ability (Laufer 1992; Albrechtsen *et al.* 2008), academic achievement (Daller *et al.* 2003), and both spoken and written communication (Wilkins 1972; Gould *et al.* 1990; Manchon *et al.* 2009). A more thorough understanding of lexical proficiency can assist L2 researchers, instructors, test developers, and educational institutions to better understand how learner lexicons develop and to make principled decisions regarding placement, pedagogy, and curricula.

Previous research has investigated lexical proficiency in a variety of ways. L2 vocabulary studies have explored the distinction between receptive and

productive knowledge (Melka 1997; Henriksen 1999; Read 2000) or the overall size of a learners' vocabulary (Meara 1992; Laufer and Nation 1999; Nation and Beglar 2007). Other areas of inquiry include the assessment of lexical performance (McNamara 1996; Polio 2001), the intrinsic difficulty of lexical items (Nation 1990; Laufer 1997), the development of L2 associative and lexical network knowledge (Haastrup and Henriksen 2000; Fitzpatrick 2006), the growth of lexical automaticity (Segalowitz *et al.* 1998; Hulstijn *et al.* 2009), and, more recently, the statistical analyses of measures of lexical output in production (Morris and Cobb 2004; Crossley *et al.* 2011).

Another promising approach to understanding lexical proficiency examines psycholinguistic measures of word knowledge (Ellis and Beaton 1993; de Groot and Keijzer 2000). This approach has provided evidence that the developing L2 lexicons are marked by the production of words that are less frequent, concrete, imageable, specific, and meaningful (Laufer and Nation 1995; Crossley *et al.* 2009; Crossley *et al.* 2010; Salsbury *et al.* 2011). However, the little work that has been done in this area focuses primarily on word frequency or judgments of word properties (e.g. human ratings of words' familiarity, imageability), neither of which tap into true on-line psycholinguistic processing features. For instance, frequency measures rely on large reference corpora as a proxy for actual language exposure, while other lexical and psycholinguistic word properties are based on subjective judgments. The aim of the current study is to determine if real-time L1 word processing information can add insight to our existing understanding of L2 lexical knowledge. To do so, we assess the productive lexical knowledge of L2 English learners in terms of word difficulty measured by psycholinguistic item response data derived from native speakers (NSs) of English. We take a multifaceted approach to operationalizing lexical proficiency, one that utilizes both cross-sectional and longitudinal data sets. Our goal is to model and predict L2 lexical proficiency in spoken language so as to better understand the development of the L2 English lexicon. We also investigate the potential for L1 psycholinguistic norms to index the intrinsic difficulty of English lexical items for L2 learners better than the most commonly used index of lexical difficulty: corpus-derived frequency measures.

LITERATURE REVIEW

Defining lexical proficiency

While L2 lexical proficiency has been operationalized in a variety of ways, the construct is traditionally approached by referring to either 'breadth' of lexical knowledge or 'depth' of understanding (Anderson and Freebody 1981). While breadth of vocabulary knowledge is typically associated with estimation of the vocabulary size of a learner (Hazenberg and Hulstijn 1996; Nurweni and Read 1999), depth of knowledge is used to refer to the quality or degree of lexical proficiency one has (Read 2000; Schmitt 2010).

Beyond the breadth-versus-depth of vocabulary distinction, a second approach to conceptualizing and assessing L2 lexical proficiency is more psycholinguistic in nature. Such an approach investigates lexical proficiency by examining the manner in which L2 lexical items are stored, processed, and retrieved from a user's mental lexicon. A key assumption among L2 researchers who study the mental lexicon is that a learner's existing lexical network undergoes restructuring as it adjusts to accommodate newly acquired words (Read 2004). As learners develop stronger links between items, it becomes easier for new words to become assimilated into the network, and the lexicon strengthens (Haastrup and Henriksen 2000).

Other than the organization of the mental lexicon as a network of formal and semantic connections, psycholinguistic researchers are also interested in how lexical items are represented and accessed in the L2 mental lexicon. One of the most attested psycholinguistic variables in predicting language processing and acquisition is word frequency (Balota and Chumbley 1990; Ellis 2002a). It is now widely accepted that frequent words are processed more quickly, both in L1 and L2 (Ellis 2002a,b), and that L2 learners typically acquire frequent words earlier than less frequent words (Read 1998; Schmitt *et al.* 2001). Other psycholinguistic variables of interest to researchers who study the L2 lexicon include concreteness, imageability, synonymy, collocation, and hypernymy (Ellis and Beaton 1993; Fitzpatrick 2006; Durrant and Schmitt 2009; Crossley *et al.* 2010; Crossley 2013). For example, studies have demonstrated that words deemed more imageable and concrete words are learned more easily (Ellis and Beaton 1993; de Groot 2006) and produced earlier (Salsbury *et al.* 2011) than less imageable and concrete words and that L2 speakers process hypernymic (i.e. hierarchical) relations in a manner similar to L1 speakers (Crossley 2013). Other studies have demonstrated that collocations are processed more slowly by L2 speakers than NSs (Siyanova and Schmitt 2008) and that collocations continue to prove difficult even for very advanced L2 learners (Nesselhauf 2005).

On-line approaches to investigating the L2 mental lexicon

By far the most prevalent approach to investigating the manner in which a learner's mental lexicon is stored, organized, and accessed involves the use of on-line psycholinguistic methods (Marinis 2010). Rather than using traditional vocabulary tests, these psycholinguistic approaches to measuring lexical proficiency are concerned with subjects' automatic processing of language and have typically employed word association, semantic priming, or word recognition tasks. By analyzing subjects' responses and reaction times (RTs), researchers are able to make inferences about L2 processing, the stages of lexical access, and the strength of network links between known words, offering insights into the organization of learners' mental lexicon (Leow *et al.* 2014). Similarly, because an efficiently organized lexical network is more accessible, studies that measure the speed with which certain lexical items are accessed

(i.e. recognized or named) allow researchers to draw conclusions about the development of L2 automaticity (DeKeyser 2001).

Researchers have also studied L2 lexical decision (LD) accuracy rates to measure the size of learners' vocabularies (Meara and Milton 2002), while others have used L2 response latencies to investigate learners' L2 processing abilities (Van Gelderen *et al.* 2004), the development of automaticity (Hulstijn *et al.* 2009), and cross-linguistic lexical representation (Kroll and Tokowicz 2001). The results from such experiments tell us about the representation of the mental lexicon and about L2 processing, specifically. However, the degree to which they inform that knowledge, particularly with regard to lexical proficiency, is still a matter of debate (Harm and Seidenberg 2004; Ratcliff *et al.* 2004; Dijkstra 2005).

Computational approaches to investigating the L2 mental lexicon

Computational and corpus linguists have begun to use natural language processing (NLP) and text analysis tools to draw conclusions about the representations of L2 learners' mental lexicon based on natural language production. Like their colleagues conducting word association or priming experiments, these researchers take a network approach to the mental lexicon. As such, computational and corpus-based approaches to investigating the mental lexicon assume a construct of lexical proficiency that is expanded to include more than the size of a learner's form-meaning associations. Instead, NLP indices are used to quantify the psycholinguistic attributes of words produced by the L2 learners, and these are in turn examined to provide evidence for learners' networks or to determine how psycholinguistic properties predict lexical proficiency.

For example, in their computational and corpus-based investigation of lexical proficiency, Crossley *et al.* (2010) evaluated the holistic ratings of L1 and L2 writing samples scored by human raters. They found that indices pertaining to word hypernymy (i.e. the number of superordinates associated with a given word where *vehicle* is a superordinate of *truck*) and frequency explained 44 per cent of the variance in human ratings of lexical proficiency. In a similar study using speaking samples, Crossley *et al.* (2011) found that indices related to word imagability, familiarity, and hypernymy predicted NS human judgments of lexical proficiency. A number of longitudinal studies have also used NLP tools to assess growth in L2 lexicons. These studies have shown that over time, L2 lexicons develop such that learners use more frequent words, less concrete words (Salsbury *et al.* 2011), less specific words (Crossley *et al.* 2009), and words with more senses (Crossley *et al.* 2010). While the bulk of computational and corpus-based studies demonstrate that more proficient learners use less frequent words, there are exceptions (Horst and Collins 2006; Crossley *et al.* 2014). For example, Crossley *et al.* (2010) reported that learners produced more frequent words over time and that the more frequent words produced

likely had more senses. Similarly, Horst and Collins found that the words learners produce increase in frequency over time while demonstrating a greater variety of basic vocabulary and a wider range of inflectional and derivational forms.

On the whole, computational approaches to lexical proficiency suggest that as L2 learners gain proficiency, they produce lexical items that are less frequent, more abstract (i.e. has fewer hypernyms, or superordinates), less specific, less concrete, and less familiar (i.e. more lexically sophisticated). Such findings are valuable in that they are derived entirely from productive lexical knowledge, versus receptive knowledge (Melka 1997; Henriksen 1999; Read 2000). Furthermore, as an assessment measure, the computational approaches described above are considered more *comprehensive* (embedded and context-dependent) than *selective* (discrete and context-independent; Read 2000; Schmitt 2010) because they investigate the naturalistic language produced by learners and amassed in learner corpora rather than responses to isolated stimuli or discrete-point tests. Overall, the investigation of L2 lexical knowledge is enhanced by NLP methods which allow researchers to automatically quantify and assess lexical phenomena in language naturally produced by learners. Still, the majority of measures available for computational analyses are derived from either large reference corpora (e.g. frequency, collocational measures) or human ratings based on subjective judgments of word properties (e.g. concreteness, imageability, familiarity). To date, no NLP measures provide information related to the automatic processing of a language by speakers in real time. To address this, the current study proposes a set of novel computational measures that bridge both on-line and computational approaches.

Current study

The current study analyzes the psycholinguistic properties of individual lexical items produced by L2 learners to investigate L2 lexical knowledge. To do so, we use a corpus-based frequency measure and four novel measures derived from L1 English speakers' item response behavior on two different psycholinguistic tasks: lexical decision (LD) and word naming (WN). Thus, the properties we analyze computationally are derived from corpora and from data gathered during on-line processing (i.e. psycholinguistic experiments of L1 English users). Our goal is to assess relations among L1 LD and WN latencies and response accuracies to predict L2 lexical proficiency, while also investigating relationships between frequency and lexical proficiency. We do so to investigate the potential for L1 behavioral norms (see 'RTs and response accuracies' section) to index the intrinsic difficulty of English lexical items beyond that of a word frequency measure. Our overarching research questions are as follows:

1 Can L1 LD and WN response latencies and accuracies be used to model L2 lexical growth in a longitudinal study, and if so, how does this compare to growth modeled by frequency?

2 Can L1 LD and WN response norms be used to explain variance in human ratings of L2 lexical proficiency beyond the variance explained by frequency?

We pursue these questions in two related studies. In Study 1, we analyze longitudinal spoken data collected from adult L2 learners over one year of study in the USA to explore if LD and WN measures predict development over time. In Study 2, we investigate the potential for LD and WN response norms to explain human ratings of lexical proficiency in L2 speech samples. In each analysis, we include a frequency measure to compare the performance of the L1 norms in explaining variance in lexical proficiency to the variance explained by frequency.

These approaches allow us to examine the validity of L1 psycholinguistic norms as predictors of lexical proficiency in L2 learners. This study builds on previous L1 research, which has used L1 response norms to examine lexical properties (e.g. Brysbaert and New's 2009 validation of a frequency measure using L1 response latencies and McDonald and Shillcock's 2001, use of L1 behavioral data to explore the construct of contextual diversity) and previous L2 research, which has used psycholinguistic word information (e.g. frequency, familiarity, imagability), but not L1 response latencies, to investigate L2 lexical proficiency (Daller *et al.* 2013; Kyle and Crossley 2015). Our method, described below, is unique in that it derives assessment measures from the psycholinguistic processing of words by NSs rather than from subjective judgments (e.g. familiarity judgments) or large reference corpora (e.g. frequency).

METHOD

RTs and response accuracies

The L1 psycholinguistic norms included in the current study were derived from two standard psychological tasks for measuring word recognition skills: LD tasks and WN tasks. LD tasks typically ask research subjects to make a forced-choice categorization of linguistic stimuli presented visually as a string of letters on a computer screen. The participant must decide if the letter string is a real word in the language of the experiment or a nonsense word. If the string is a real word, the participant presses a key assigned to the identification of real words; if the string is not a real word in the language of the experiment, a different key is pressed. Both the time it takes for the participant to press a key (RT) and the error rate (accuracy) are measured.

In the current study, RTs for lexical items measured the amount of time it took (in milliseconds) for participants to determine that written words were in fact real words in English, while response accuracies measured participants' ability to distinguish correctly between real words and pseudowords (i.e. nonwords that are plausible given the rules of word formation in a particular language). While correct rejections of nonsense words in the stimuli did

factor into participants' rate of accuracy, these RTs were not included in analysis because the current study was only interested in RTs to real words.

Another common psychological tool used to measure word recognition skills is the WN task. In WN tasks, participants are presented with an orthographic word that they then have to name (i.e. read) aloud. The time it takes for the participant to begin pronouncing the word is measured as the RT. Error rates are calculated as well.

The RTs and accuracies used in the current article were derived from the LD and WN tasks conducted for the English Lexicon Project (ELP), a large publicly available behavioral and descriptive data set (Balota *et al.* 2007). This data set includes LD and WN response norms from 816 participants, with approximately 3,400 responses per participant for the LD task and 2,500 per participant for the WN tasks. RTs, their standard deviations (SDs), and accuracies were calculated in response to 80,962 real-word and nonword stimuli (40,481 in each condition), including both mono- and multi-syllabic words. All participants were native English speakers (NESs) recruited from participant pools across six different universities in the USA. Real words were selected to represent a range of frequencies (based on Kučera and Francis 1967 and CELEX norms; Baayen *et al.* 1993) while the nonwords were generated by changing one to two letters in selected real words.

Average RTs, SDs, and accuracies for each word were then coded into the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle and Crossley 2015), a freely available computational tool that calculates hundreds of automated indices related to lexical sophistication. TAALES LD and WN indices were calculated by taking the sum of the LD and WN values for all the words in a speech sample (see below for a description of the corpora from which the speech samples were taken) and dividing that sum by the total number of words in the sample. Accuracies represent the proportion of accurate responses for a particular word on a given task. If a word in a given speech sample was not represented in the LD or WN list, the word was not included in the calculation of the index. However, the likelihood of speakers producing a word that was not represented was very small because the ELP norms provide values for over 40,000 English words.

Frequency measure

The frequency measure used in the current study is derived from the 8,388 American film and television subtitles comprising the 51 million-word $SUBTLEX_{US}$ corpus (Brysbaert and New 2009). The index reports the mean $SUBTLEX_{US}$ log frequency score for all words in a given text. This measure is unique from many other corpus-based frequencies because it is based on entirely spoken (though not entirely naturalistic) data, making it more suitable for the analysis of spoken data. Furthermore, $SUBTLEX_{US}$ frequencies have been shown to be more predictive of human processing latencies for words

used in psycholinguistic research than frequency indices (Brysbaert and New 2009). The frequency data used in this study was reported by TAALES.

STUDY 1

In Study 1, we analyze longitudinal spoken L2 data to assess the ability of L1 psycholinguistic latencies and accuracies to model lexical development over time. We compare our results to the linear development of frequency in the same data. The longitudinal spoken corpus used in Study 1 contains unrehearsed spoken language collected from six L2 English learners bimonthly for one year (see Crossley *et al.* 2010 and further description below).

Study 1 asks the following research questions:

1 Do the L1 LD and WN response measures derived from lexical items produced by English L2 learners exhibit a significant linear trend as learners spend more time in an L2 environment?
2 If so, what is the strength of the relationship of these trends relative to the linear development of frequency in the same data?

Method

Corpus

The corpus analyzed in Study 1 was derived from a longitudinal study by Crossley *et al.* (2010). The corpus contained unprepared spoken language collected from L2 English learners bimonthly (i.e. once every two weeks) for one year. All learners were enrolled in an intensive English program at a large American university and had lived in the USA for three weeks or less prior to data collection. Upon arrival to the program, the learners placed into the lowest proficiency level of the program (i.e. Level 1 of a six-level program). Placement was based upon an internal assessment instrument. Four L1 language backgrounds were represented by participants: Spanish ($n = 1$), Japanese ($n = 1$), Korean ($n = 1$), and Arabic ($n = 3$).

Throughout the course of a year, learners participated in regular elicitation sessions with NES interviewers. Learners responded to 50 unique prompt topics during these sessions and added their own topics as well. Elicitation sessions were balanced in terms of prompt topics presented; however, much of the data was composed of naturalistic discourse that was not topic-based. All spoken interactions were transcribed for analysis. Only learners' language (vs. the language of interlocutors) was included in the current analysis. Beginning at the end of their second month of study, all participants were assessed through the institutional Test of English as a Foreign Language (TOEFL) examination, which includes reading, grammar, and vocabulary (but no speaking) sections. The TOEFL exam was administered every other month thereafter to assess learners' language growth. Follow-up analysis demonstrated a significant increase in

learners' average TOEFL scores as a function of time (see Crossley *et al.* 2010), suggesting that learners indeed experienced overall linguistic growth during the year in which this corpus was collected.

Data analysis

Indices measuring each transcribed speech sample's frequency, ELP RTs, and ELP accuracies were reported by TAALES (Kyle and Crossley 2015). For frequency, a TAALES index reported the average $SUBTLEX_{US}$ log frequency for all words produced in a given speech sample. RT and accuracy values represent the average value for each word in the sample text that was included in the ELP database. To best capture longitudinal development, only those speech samples collected during mostly equivalent intervals (i.e. Weeks 2, 11, 21, 31, 45, and 49) were included in the current analysis (Times 1, 2, 3, 4, 5, and 6 respectively, see Table 1). Each speech sample analyzed was collected roughly 1.5 months after the previous sample.

Prior to running inferential statistics, a visual examination of the data indicated that the assumptions for parametric statistics had been met. Repeated-measures analyses of variance were then conducted to look for significant linear trends in frequency, LD, and WN variables between Times 1 and 6.

Results

Study 1 tested the hypothesis that ELP measures and frequency would demonstrate significant change over time as L2 learners developed in lexical proficiency. Descriptive statistics were calculated for the ELP measures and frequency of the learners' output across six data collection times of interest (Table 2).

Significant linear trends were reported for our selected frequency index, $F(1, 5) = 30.866$, $p = .003$, partial-$\eta^2 = .861$); LD accuracy index, $F(1, 5) = 6.643$, $p = .050$, partial-$\eta^2 = .571$); WN RT index, $F(1, 5) = 8.167$, $p = .035$, partial-$\eta^2 = .620$); and WN accuracy index, $F(1, 5) = 12.883$, $p = .016$, partial-$\eta^2 = .720$). The effect sizes reported here (partial-η^2) indicate the variance

Table 1: Study 1 speech sample collection times

Time	Week	Dates of collection
Time 1	Week 2	9–10 September
Time 2	Week 11	11–13 November
Time 3	Week 21	21–22 January
Time 4	Week 31	31 March–3 April
Time 5	Week 45	8–17 July
Time 6	Week 49	3–7 August

Table 2: Average frequency and ELP measures for longitudinal spoken data

Time	SUBTLEXUS mean frequency (logged)	LD task		WN task	
		Mean RT	Mean accuracy	Mean RT	Mean accuracy
1	4.667	611.315	0.971	598.853	0.994
2	4.877	608.550	0.968	596.720	0.996
3	4.795	611.914	0.968	597.235	0.994
4	4.984	608.781	0.968	594.032	0.994
5	4.970	611.082	0.968	596.239	0.993
6	4.993	612.991	0.967	595.546	0.993

explained by a given predictor. As a general rule, a partial eta-squared of .14 or higher is considered a large effect size (Cohen 1988). The results indicate that—contrary to expectations but in line with Crossley *et al.* (2010)—L2 learners produce more frequent words as a function of time (however, see our caveat regarding this finding in the discussion below). In contrast, L2 learners produced words with lower LD accuracy, lower WN RT, and lower WN accuracy (i.e. L2 learners produced more sophisticated words in terms of on-line processing metrics over time). With regard to the strength of the ELP measures relative to word frequency, the frequency index used here demonstrated the strongest linear trend (partial-$\eta^2 = .861$), followed by WN accuracy (partial-$\eta^2 = .720$), WN RTs (partial-$\eta^2 = .620$), and LD accuracy (partial-$\eta^2 = .571$).

STUDY 2

Whereas Study 1 analyzed the development of the lexicon over time, Study 2 uses L1 psycholinguistic norms (from the WN and LD tasks) and a frequency measure to explain human ratings of lexical proficiency in a cross-sectional corpus of L2 speech samples. Study 2 asks the following research question:

1 Do L1 LD and WN response norms predict variance in human ratings of lexical proficiency beyond variance predicted by frequency?

Method

Corpus

The corpus used in Study 2 comprises transcribed spoken data collected from both L2 English learners and NESs (Crossley and Salsbury 2011). The L2 English learner participants were from two different universities in the USA. Participants were enrolled either as undergraduates or as students in an intensive English program. L1 backgrounds of the learners included Korean, Arabic,

Mandarin, Spanish, French, Japanese, and Turkish. Transcribed data were collected during naturalistic, interactional spoken discourse between a non-NES (i.e. the learner) and a NES interlocutor.

To ensure variance across speech samples, prior to data collection, learners were categorized into beginning, intermediate, and advanced proficiency levels based on the TOEFL or ACT ESL Compass scores (see Crossley and Salsbury 2011, for details).[1] These levels were used to guarantee a range of proficiency variance in the speech samples that comprise the corpus; however, the levels themselves were not used as dependent variables. Sixty samples of L2 speech were collected from each proficiency level, resulting in 180 L2 speech samples in total. An additional 60 NES speech samples were selected from the *Switchboard* corpus (Godfrey and Holliman 1993), a collection of approximately 2,400 telephone conversations taken from 543 speakers from all areas of the USA. The conversations are all naturalistic, involve two interlocutors, and cover a range of topics. In total, the speech sample corpus for this study included 240 speech samples (180 L2 speech samples and 60 NES samples).

Human ratings

Human raters were three graduate students from an English department at a large university in the USA. All three raters were functionally monolingual NESs. Ratings of speech samples were based on transcriptions[2] that contained the complete interactions between the speaker of interest and the interlocutor. Samples of the speaker of interest were controlled for word length by selecting segments of the speech that contained approximately 150 words. Each segment was collected so as to avoid artificially separating utterances. The three trained raters assessed the 240 speech samples using a holistic grading rubric adapted from the 'American Council on the Teaching of Foreign Languages' (ACTFL) proficiency guidelines for speaking and writing (ACTFL 1999) and proficiency rubrics produced by American College Testing (ACT) and the College Board (for further details, see Crossley and Salsbury 2011). Raters were asked to assess lexical proficiency using a 5-point Likert scale. A score of 5 was used to indicate high lexical proficiency while a score of 1 was used to identify samples that demonstrated little lexical mastery (see Appendix A and Crossley and Salsbury 2011 for further details). Final holistic lexical proficiency scores used for analysis were an average of the three raters. The average correlation between the three raters was $r = .796$ ($p < .001$) with a weighted correlation of $r = .921$.

Frequency and ELP measures

While human raters based their proficiency ratings on speech samples that included the utterances of both the learner and his/her interlocutor, following Crossley *et al.* (2011), the samples used for computational and statistical analysis in the current study contained only the utterances of the learner.

Data analysis

The holistic human ratings of spoken data were used as the dependent variable in a multiple regression analysis with the following independent variables: $SUBTLEX_{US}$ frequency, LD item mean RTs, LD item mean accuracy, WN item mean RTs, and WN item mean accuracy.

We first conducted correlations between holistic ratings and the predictor variables described above to assess which measures reported a meaningful ($r \geq .10$ indicating at least a small effect size) and significant relationship ($p < .05$) with lexical proficiency. Correlations among the variables that demonstrated a meaningful and significant relationship were then checked for multicollinearity. If any two variables were highly collinear ($r > .90$), only the variable with the strongest relationship to holistic ratings was retained. The remaining variables were entered as independent variables into a stepwise multiple regression to explain the variance in holistic ratings each speech segment received.

Prior to carrying out the regression analysis, we divided speech samples into training and test sets using a 67/33 split (approximately 67 per cent training, 33 per cent test) (Witten *et al.* 2011), which allowed for cross-validation of the regression model. The assumption of this cross-validation procedure is that if the model derived from a training set predicts the outcome variable in the test set at a similar accuracy rate as the training set, the regression model can be considered stable. We first obtained a model from the speech samples comprising the training set (approximately 67 per cent of the spoken samples). We then applied that model to the test set (approximately 33 per cent of samples) to assess its predictive power and overall generalizability.

Results

To determine which measures would best predict holistic ratings, we verified that each of the five variables demonstrated a significant correlation with holistic ratings ($r \geq .10$, $p < .05$). Correlations between the five variables and holistic ratings are displayed in Table 3. Four of five variables demonstrated significant correlations: LD RT, WN accuracy, WN RT, and frequency. The correlations for the on-line indices were in line with predictions. However, the correlations for the frequency index were positive indicating that more frequent words were indicative of lexical proficiency, which was not in line with predictions. In addition, the on-line indices demonstrated stronger correlations with lexical proficiency than the frequency index. These five indices were then checked for multicollinearity. None of these were highly collinear ($r > .90$).

A step-wise regression was conducted to determine the strength of the four indices in predicting holistic human ratings. An initial model yielded a significant result with four indices, $F(4, 148) = 16.865$, $p < .001$, $r = .560$, $R^2 = .313$. The model demonstrated that LD RTs, frequency, WN RTs, and WN accuracy

Table 3: Correlations between holistic human ratings and measures of interest

Variable	Holistic rating	
	r	*p*
LD mean RT	0.362	0
WN mean accuracy	−0.344	0
WN mean RT	0.179	0.005
$SUBTLEX_{US}$ mean frequency (logged)	0.150	0.020
LD mean accuracy	−0.102	0.115

Table 4: Indices regressed on holistic human ratings training set model

Entry	Predictors included	*R*	R^2	R^2 change	β	*SE*	*B*	*t*	*p*
1	LD mean RT	0.370	0.137	0.137	0.273	0.011	0.036	3.312	0.001
2	$SUBTLEX_{US}$ mean frequency (logged)	0.453	0.205	0.068	0.443	0.516	2.682	5.196	0.000
3	WN mean RT	0.524	0.275	0.07	0.311	0.017	0.056	3.294	0.001
4	WN mean accuracy	0.560	0.313	0.038	−0.219	28.929	−83.202	−2.876	0.005

Note: Estimated constant term is 17.356; β is standardized coefficient; *SE* is standard error; *B* is unstandardized coefficient.

explained approximately 31 per cent of the variance in the holistic lexical proficiency ratings of the speech samples (Table 4). When the same model was applied to the test set, the model yielded $r = .507$, $R^2 = .257$, indicating that the same measures explained roughly 26 per cent of the variance in the human ratings for the test set and demonstrating that the model was stable. The standardized coefficients (β) in Table 4 indicate the number of SD changes we would expect in lexical proficiency for a one SD change in any given index. Table 4 indicates that the strongest increases occur for LD RT (i.e. for every one SD increase in average LD RT value for a given speech sample, we can expect to see a 0.273 SD increase in lexical proficiency ratings) followed by our frequency index, WN RT, and WN accuracies.

DISCUSSION

Studies that have examined L2 lexical proficiency using computational methods have traditionally investigated features related to frequency, imageability, concreteness, meaningfulness, and range (Morris and Cobb 2004; Crossley *et al.* 2010; Crossley *et al.* 2011; Kyle and Crossley 2015). These studies have found that L2 learners exhibit growth in lexical proficiency in terms of the production of less frequent words, greater lexical diversity, and the use of less

imageable, concrete, and meaningful language (Ellis and Beaton 1993; Crossley *et al.* 2010; Salsbury *et al.* 2011; Crossley *et al.* 2013).

The purpose of the studies reported in this article was to extend and expand our knowledge of L2 lexical proficiency by examining new indices derived from L1 psycholinguistic item response data. The two studies reported here demonstrate that L1 psycholinguistic norms can be used to predict L2 lexical proficiency and lexical growth across different learner corpora. In all cases, L1 psycholinguistic norms followed predicted patterns while our frequency index did not. In the case of our cross-sectional data set (Study 2), the results also suggest that an ELP measure (LD RTs) explained more variance in holistic lexical proficiency than a word frequency measure. These findings are important because they suggest that on-line psycholinguistic data derived from L1 subjects can add unique information to our understanding of L2 lexical knowledge. The findings are distinct from other investigations of psycholinguistic word properties because the ELP indices analyzed here do not originate from reference corpora or human judgments of word properties but from the on-line processing of words during actual psycholinguistic tasks. The two studies reported here contribute further to the literature in that they include analysis of spoken learner corpora that include both cross-sectional and longitudinal data. Below, we discuss the results of the individual studies and provide implications for their importance in SLA and theories of lexical organization.

Study 1

The first research question for Study 1 asked whether L1 LD and WN measures derived from the lexical items produced by English L2 learners exhibited significant linear trends as learners spent more time in an L2 environment. Results demonstrated that LD accuracies, WN RTs, and WN accuracies demonstrated significant linear trends over time, with statistically large effect sizes. These findings suggest words that resulted in longer WN RTs and lower WN and LD accuracies in experimental data were more likely to be produced later (and at higher proficiency levels) by L2 learners. Overall, these results demonstrate that as learners develop lexical proficiency over time, the words they produce are named more slowly and inaccurately and more often incorrectly judged as nonsense words by L1 speakers in experimental settings.

The second research question explored by Study 1 compared the relative strength of the linear trend demonstrated by ELP measures in learners' lexicon over time to a corpus-derived frequency measure. Results indicate that the $SUBTLEX_{US}$ frequency index demonstrated the strongest linear trend, with an effect size (prtl-η^2) of .861. The next strongest linear trend was found in the WN accuracy measure (prtl-$\eta^2 = .720$), followed by WN response times (prtl-$\eta^2 = .620$) and LD accuracies (prtl-$\eta^2 = .571$). Interestingly, the linear trend demonstrated by frequency runs counter to what might be predicted from past research: As the learners analyzed in this study gained exposure to English, they actually produced more frequent words. A similar finding was

demonstrated by Crossley *et al.* (2010), who found that the use of more frequent words over time by learners correlated with the use of more polysemous words, suggesting that the more frequent words may have had more senses. In another study, Crossley *et al.* (2014) demonstrated that the apparent production of more frequent words over time in L2 data was due to repetition of infrequent words in learner speech during early stages of language acquisition, which may be the case in this study as well. For example, at Time 1, one learner repeats the infrequent word *psychology* nine times in one turn while explaining his/her area of study. Another learner repeats the infrequent word *aerobics* five times in what appears to be an unsuccessful attempt to recall another word that better describes his/her preference for exercise.

Study 2

Whereas Study 1 examined changes in learners' lexical production over the course of one year, Study 2 analyzed human ratings of lexical proficiency cross-sectionally. Our research question for Study 2 asked whether L1 psycholinguistic measures could be used to predict human ratings of lexical proficiency in a corpus of L2 speech samples. A frequency measure was included in our model to determine the predictive ability of the ELP measures beyond frequency. Our regression analysis (Table 4) found that 31 per cent of the variance in holistic ratings of lexical proficiency was explained by a combination of LD RTs, frequency, WN RTs, and WN accuracy measures derived from the lexical items produced in each speech sample.

LD RTs correlated positively with human ratings, meaning that words with longer average RTs in the L1 ELP data set were produced more often by L2 speakers rated as more lexically proficient. Thus, speech samples with words that took longer to identify as English words by L1 speakers were identified as more lexically proficient. WN accuracy correlated negatively with holistic ratings. This finding suggests that speech samples that contained words that were more difficult to name aloud in the L1 LD task were more likely to be rated as more lexically proficient. In addition, WN RTs correlated positively with lexical proficiency, meaning that more proficient speakers produced words that are read aloud more slowly by NESs. Frequency also correlated positively with holistic ratings of lexical proficiency, meaning that speakers rated as more proficient actually used more frequent words. Like Study 1, this finding runs counter to the trend observed in previous research. A post hoc analysis of the data indicates that this finding may be the result of measuring the frequency of all words as compared to content words alone. When the data were re-analyzed using the same frequency measures for function words and content words, the correlation for function words demonstrated a small and positive correlation ($r = .146$, $p = .043$) while the correlation for content words demonstrated a small, but negative correlation ($r = -.196$, $p = .002$), as would be predicted. Both correlations, however, were weaker than those reported by LD RT and WN accuracy.

Overall, Study 2 demonstrates that speech samples rated as more lexically proficient contained words that are on average recognized more slowly, and named more slowly and with less accuracy by L1 speakers. Combined, the three ELP indices in our model seem to capture a word's intrinsic difficulty. It may be that these measures index L2 lexical proficiency because a word's intrinsic difficulty—as determined by L1 LD and WN behavioral data—impacts its likelihood of being acquired by an L2 speaker. Results also indicate that LD RTs explained *more* variance in holistic ratings of lexical proficiency (13.7 per cent) than the $SUBTLEX_{US}$ frequency measure (6.8 per cent) (see Table 4). Meanwhile, WN RTs and accuracies explained 7 and 3.8 per cent of variance in lexical proficiency, respectively (see Table 4).

Overarching research questions

The core question motivating this research was whether L1 psycholinguistic norms could be used to predict L2 spoken lexical proficiency. Study 1 took a longitudinal approach to lexical growth, investigating the language produced by learners over the course of an entire year. Study 2 operationalized L2 lexical proficiency using holistic human ratings derived from transcriptions of spoken language collected at a single point in time. Results from both studies demonstrate that LD and WN response data can indeed be used to predict a portion of L2 spoken lexical proficiency, regardless of the length of the study or the manner of operationalizing lexical proficiency. While frequency exhibited the strongest linear trends in longitudinal data, three ELP measures also developed linearly and demonstrated strong effect sizes. Furthermore, frequency trends over time were not in the expected direction, suggesting that the role of frequency in predicting lexical growth warrants further investigation. In addition, Study 2 found that LD RTs explained more variance in holistic lexical proficiency ratings than a frequency measure, while frequency and WN RT were roughly comparable. These findings provide tentative evidence for the strength of certain psycholinguistic norms relative to frequency data in explaining L2 lexical proficiency; however, the relationship between L1 norms, frequency, and proficiency should certainly be explored further.

Taken as a whole, this research demonstrates that the on-line processing of words by NSs can contribute to our understanding of L2 lexical production and acquisition as well as to a words' intrinsic difficulty for L2 learners. Furthermore, if L1 psycholinguistic measures like the ones used here can be used to index L2 lexical proficiency, language assessments, pedagogical choices, and curricular decisions could be informed through human processing times and accuracies as compared to measures derived from corpora or subjective judgments.

CONCLUSION

We assessed the productive lexical knowledge of L2 English learners using psycholinguistic item response data derived from NESs. Our results demonstrated that L1 psycholinguistic RTs and accuracies can be used to index L2 spoken lexical proficiency and lexical growth across different learner corpora. Study 1, which analyzed longitudinal data, indicated that as L2 learners develop proficiency over time, they produce words that are recognized with less accuracy (higher LD accuracy) and named more slowly (longer WN response times) and with less accuracy (lower WN accuracy) by L1 speakers in experimental settings. Study 1 also demonstrated the linear development of frequency over time, with learners producing more frequent words as they spent more time in an English-speaking environment. Study 2, a cross-sectional study, revealed that learners rated as more lexically proficient by human raters produced words that were recognized more slowly (longer LD response latencies) and named more slowly (longer WN response times) and with less accuracy (lower WN accuracy) by L1 speakers. Study 2 also demonstrated that more lexically proficient speakers used more frequent words, though the amount of variance explained by one ELP variable (LD response times) was higher than the variance explained by frequency. Taken together, these two studies demonstrate that on-line L1 psycholinguistic information can indeed add unique information to our understanding of L2 lexical knowledge, with L1 norms offering greater explanatory power independent from frequency in cross-sectional data.

The limitations of the current study present avenues for future research. For example, measures like the type used in this study are problematic because they treat the lexicon as consisting solely of single-word units and ignore the reality of phraseological items (Römer 2009). This assumption that a word in isolation is the base unit of language ignores much contemporary research that emphasizes the inseparability of lexis and grammar and the existence of meaning at different levels of complexity (Wray 2002; Römer 2009). Future studies of this nature would do well to include analysis of multiword units as they develop with increased proficiency (Crossley and Salsbury 2011).

The findings also raise issues with regard to frequency that should be addressed in future studies of this nature. Specifically, our results suggest that a more appropriate measure of frequency in learner speech may be one that calculates frequency for the types produced within a sample, rather than through frequency tokens. In this manner, the frequency of a repeated word would only contribute to the average one time. Similarly, these findings demonstrate the need for further refinement of the LD and WN indices to distinguish between content and function words.

While our results suggest that L1 psycholinguistic response data can help explain variance in L2 learners' lexical proficiency beyond frequency, we certainly have not explained all variance. In particular, our method of analysis did not allow us to capture productive elements of spoken language, such as pronunciation, prosody, and verbal fluency. Clearly, more work is needed to

adequately describe lexical proficiency, particularly in spoken data. Also, the learner sample sizes in both studies were quite small. A larger data set, with more consistent and frequent data collection intervals, would be beneficial in future studies that replicate our method.

Further investigation is also needed to examine the relationship between L1 response norms and psycholinguistic word properties beyond frequency (e.g. concreteness, familiarity, imagability), as they relate to lexical proficiency and L2 learner production. Finally, while the two studies presented here have demonstrated the L1 psycholinguistic RTs and accuracies in the ELP data set can be used to gain insight into L2 lexical proficiency, there is clear potential for a parallel psycholinguistic data set derived from L2 speakers' behavioral data.

NOTES

1 Institutional TOEFL and ACT ESL Compass reading scores were used to classify L2 writers into beginning, intermediate, and advanced categories. For TOEFL, we used the total score on the exam and referred to Wendt and Woo (2009) and Boldt *et al.* (1992) in establishing our three broad proficiency categories. However, no comparisons are available between the TOEFL tests and the ACT ESL Compass test. Thus, we relied on the test makers' suggested proficiency levels and descriptors in classifying the ACT group of students. The classification of the L2 writers is as follows: L2 writers who scored 400 or below on the TOEFL or 126 or below on the combined Compass ESL reading/grammar tests were classified as beginning level. L2 writers who scored between 401 and 499 on the TOEFL PBT or 127 and 162 on the combined Compass ESL reading/grammar tests were classified as intermediate level. L2 writers who scored 500 or above on the TOEFL PBT or 163 or above on the combined Compass ESL reading/grammar tests were classified as advanced level. If TOEFL scores and ACT ESL Compass reading scores were available, we relied on TOEFL scores. If only ACT ESL Compass reading scores were available, we relied on those scores alone.

2 Transcriptions for analysis were modified to exclude interjections (e.g. 'uhm') as well as any non-English or invented words. Most proper nouns were left in the data. Non-target-like forms of the irregular past tense (e.g. 'sleeped') were left un-changed, as were reductions such as 'gonna'.

REFERENCES

Albrechtsen, D., **K. Haastrup**, and **B. Henriksen**. 2008. *Vocabulary and Writing in a First and Second Language*. Palgrave Macmilan.

Alderson, J. C. 2005. *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. Continuum.

American Council on the Teaching of Foreign Languages. 1999. 'ACTFL proficiency guidelines: Speaking,' available at: www.actfl.org/files/public/Guidelinesspeak.pdf, Accessed 13 December 2008.

Anderson, J. R. and **P. Freebody**. 1981. 'Vocabulary knowledge' in J. Guthrie (ed.): *Comprehension and Teaching: Research Reviews*. International Reading Association, pp. 77–117.

Baayen, R. H., **R. Piepenbrock**, and **H. van Rijn**. 1993. *The CELEX Lexical Database.*

Linguistic Data Consortium. University of Pennsylvania.

Balota, D. A. and **J. I. Chumbley**. 1990. 'What are the effects of frequency in visual word recognition tasks? Right where we said they were!,' *Journal of Experimental Psychology: General* 133: 231–7

Balota, D. A., **J. Yap**, **K. A. Hutchison**, **M. J. Cortese**, **B. Kessler**, **B. Loftis**, **A. H. Neely**, **D. L. Nelson**, **G. B. Simpson**,and **R. Treiman**. 2007. 'The English Lexicon Project,' *Behavior Research Methods* 39/3: 445–59.

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

Boldt, R. F., **D. Larsen-Freeman**, **M. S. Reed**, and **R. G. Courtney**. 1992. 'Distributions of ACTFL Ratings by TOEFL Score Ranges,' TOEFL Research Report No. 41. Educational Testing Service.

Brysbaert, M. and **B. New**. 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* 41/4: 977–90.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn.. Lawrence Erlbaum Associates.

Crossley, S., **T. Salsbury**, **A. Titak**, and **D. McNamara**. 2014. 'Frequency effects and second language lexical acquisition: Word types, word tokens, and word production,' *International Journal of Corpus Linguistics* 19/3: 301–32.

Crossley, S. A. 2013. 'Assessing automatic processing of hypernymic relations in first language speakers and advanced second language learners: A semantic priming approach,' *The Mental Lexicon* 8/1: 96–116.

Crossley, S. A., **T. Cobb**, and **D. S. McNamara**. 2013. 'Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications,' *System* 41/4: 965–81.

Crossley, S. A., **T. Salsbury**, and **D. McNamara**. 2009. 'Measuring L2 lexical growth using hypernymic relationships,' *Language Learning* 59/2: 307–34.

Crossley, S. A., **T. Salsbury**, and **D. McNamara**. 2010. 'The development of polysemy and frequency use in English second language speakers,' *Language Learning* 60/3: 573–605.

Crossley, S. A., **T. Salsbury**, **D. S. McNamara**, and **S. Jarvis**. 2010. 'Predicting lexical proficiency in language learner texts using computational indices,' *Language Testing* 28/4: 561–80.

Crossley, S. A., **T. Salsbury**, **D. S. McNamara**, and **S. Jarvis**. 2011. 'What is lexical proficiency? Some answers from computational models of speech data,' *TESOL Quarterly* 45/1: 182–93.

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

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

Daller, M., **J. Turlik**, and **I. Weir**. 2013. 'Vocabulary acquisition and the learning curve' in J. S. Daller and M. Daller (eds): *Vocabulary Knowledge: Human Ratings and Automated Measures*, Vol. 47. John Benjamins, pp. 187–217.

David, A. 2008. 'A developmental perspective on productive lexical knowledge in L2 oral interlanguage,' *Journal of French Language Studies* 18/3: 315–31.

de Groot, A. 2006. 'Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting,' *Language Learning* 56/3: 463–506.

de Groot, A. and **R. Keijzer**. 2000. 'What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting,' *Language Learning* 50/1: 1–56.

DeKeyser, R. M. 2001. 'Automaticity and automatization' in P. Robinson (ed.): *Cognition and Second Language Instruction*. Cambridge University Press, pp. 125–51.

Dijkstra, T. 2005. 'Bilingual visual word recognition and lexical access' in F. F. Kroll and A. M. B. De Groot (eds): *Handbook of Bilingualism: Psycholinguistics Approaches*, Oxford University Press, pp. 179–201.

Durrant, P. and **N. Schmitt**. 2009. 'To what extent do native and non-native writers make use of collocations?,' *IRAL-International Review of Applied Linguistics in Language Teaching* 47/2: 157–77

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

Ellis, N. C. 2002b. 'Reflections on frequency effects in language processing,' *Second Language Acquisition* 24: 297–339.

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

Fitzpatrick, T. 2006. 'Habits and rabbits: Word associations and the L2 lexicon,' *EUROSLA Yearbook* 6: 121–45.

Godfrey, J. J. and **E. Holliman**. 1993. *Switchboard-1 [C-D1-ROMJ*. Linguistic Data Consortium.

Gould, R., **P. Nation**, and **J. Read**. 1990. 'How large can a receptive vocabulary be?,' *Applied Linguistics* 11/4: 341–63

Haastrup, K. and **B. Henriksen**. 2000. 'Vocabulary acquisition: Acquiring depth of knowledge through network building,' *International Journal of Applied Linguistics* 10/2: 221–40.

Harm, M. W. and **M. S. Seidenberg**. 2004. 'Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes,' *Psychological Review* 111/3: 662–720.

Hazenberg, S. and **J. H. Hulstijn**. 1996. 'Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation,' *Applied Linguistics* 17/2: 145–63.

Henriksen, B. 1999. 'Three dimensions of vocabulary development,' *Studies in Second Language Acquisition* 21/2: 303–17.

Horst, M. and **L. Collins**. 2006. 'From "faible" to strong: How does their vocabulary grow? *The,' Canadian Modern Language Review* 63/1: 83–106 /

Hulstijn, J. H., **A. Van Gelderen**, and **R. Schoonen**. 2009. 'Automatization in second language acquisition: What does the coefficient of variation tell us?,' *Applied Psycholinguistics* 30/4: 555–82

Kroll, J. F. and **N. Tokowicz**. 2001. 'The development of conceptual representation for words in a second language' in J. L. Nicol and T. Langendoen (eds): *Language Processing in Bilinguals*. Blackwell, pp. 49–71.

Kučera, H. and **W. Francis**. 1967. *Computational Analysis of Present-day American English*. Brown University Press.

Kyle, K. and **S. A. Crossley**. 2015. 'Automatically assessing lexical sophistication: Indices, tools, findings, and application,' *TESOL Quarterly* 49/4: 757–86.

Laufer, B. 1992. 'Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability?,' *Journal of Research in Reading* 15/2: 95–103

Laufer, B. 1997. 'What's in a word that makes it hard or easy? Intralexcial factors affecting the difficulty of vocabulary acquisition' in N. Schmitt and M. McCarthy (eds): *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge University Press, pp. 140–55.

Laufer, B. and **P. Nation**. 1995. 'Vocabulary size and use: Lexical richness in L2 written production,' *Applied Linguistics* 16/3: 307–32.

Laufer, B. and **P. Nation**. 1999. 'A vocabulary-size test of controlled productive ability,' *Language Testing* 16/1: 33–51.

Leow, R. P., **S. Grey**, **S. Marijuan**, and **C. Moorman**. 2014. 'Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview,' *Second Language Research* 30/2: 111–27.

Manchon, R. M., **L. Murphy**, and **J. R. de Larios**. 2009. 'Lexical retrieval processes and strategies in second language writing: A synthesis of empirical research,' *International Journal of English Studies* 7/2: 149–74.

Marinis, T. 2010. 'Using on-line processing methods in language acquisition research' in E. Blom and S. Unsworth (eds): *Experimental Methods in Language Acquisition Research*. John Benjamins, pp. 139–62.

McDonald, S. A. and **R. C. Schillcock**. 2001. 'Rethinking the word frequency effect: The neglected role of distributional information in lexical processing,' *Language and Speech* 44/3: 295–323.

McNamara, T. F. 1996. *Measuring Second Language Performance*. Addison Wesley Longman.

Meara, P. 1992. *EFL Vocabulary Tests*. Centre for Applied Language Studies.

Meara, P. and **J. L. Milton**. 2002. *X_Lex: The Swansea Vocabulary Levels Test*. Express.

Melka, F. 1997. 'Receptive vs. productive aspects of vocabulary' in N. Schmitt and M. McCarthy (eds): *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge University Press, pp. 84–102.

Morris, L. and **T. Cobb**. 2004. 'Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees,' *System* 32/1: 75–87.

Nation, I. S. P. 1990. *Teaching and Learning Vocabulary*. Heinle and Heinle Publishers.

Nation, I. S. P. and **B. Beglar**. 2007. 'A vocabulary size test,' *The Language Teacher* 31/7: 9–13.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*, Vol. 14. John Benjamins Publishing.

Nurweni, A. and **J. Read**. 1999. 'The English vocabulary knowledge of Indonesian university students,' *English for Specific Purposes* 18/2: 161–75.

Polio, C. 2001. 'Research methodology in second language writing research: The case text-based studies' in T. Silva and P. K. Matsuda (eds): *On Second Language Writing*. Routledge, pp. 91–115.

Ratcliff, R., **P. Gomez**, and **G. McKoon**. 2004. 'A diffusion model account of the lexical decision task,' *Psychological Review* 111/1: 159–82.

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

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

Read, J. 2004. 'Plumbing the depths: How should the construct of vocabulary knowledge be defined?' in P. Bogaards and B. Laufer (eds): *Vocabulary in a Second Language*. John Benjamins Publishing, pp. 209–28.

Römer, U. 2009. 'The inseparability of lexis and grammar: Corpus linguistic perspectives,' *Annual Review of Cognitive Linguistics* 7/1: 140–162.

Salsbury, T., **S. A. Crossley**, and **D. S. McNamara**. 2011. 'Psycholinguistic word information in second language oral discourse,' *Second Language Research* 27/3: 343–60.

Schmitt, N. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave Macmillan.

Schmitt, N., **D. Schmitt**, and **C. Clapham**. 2001. 'Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test,' *Language Testing* 18/1: 55–88.

Segalowitz, S. J., **S. Segalowitz**, and **A. G. Wood**. 1998. 'Assessing the development of automaticity in second language word recognition,' *Applied Psycholinguistics* 19/1: 53–67.

Siyanova, A. and **N. Schmitt**. 2008. 'L2 learner production and processing of collocation: A multi-study perspective,' *Canadian Modern Language Review* 64/3: 429–58.

Van Gelderen, A., **R. Schoonen**, **K. De Glopper**, **J. Hulstijn**, **A. Simis**, **P. Snellings**, and **M. Stevenson**. 2004. 'Linguistic knowledge, processing speed, and metacognitive knowledge in first-and second-language reading comprehension: A componential analysis,' *Journal of Educational Psychology* 96: 1–19.

Wendt, A. and **A. Woo**. 2009. 'A minimum english proficiency standard for the test of English as a foreign language internet-based test (TOEFL-iBT)' in *NCLEX Psychometric Research Brief*. National Council of State Boards of Nursing.

Wilkins, D. 1972. 'Do reading and interactive vocabulary instruction make a difference? An empirical study,' *TESOL Quarterly* 31/1: 121–40.

Witten, I. A., **E. Frank**, and **M. A. Hall**. 2011. *Data mining: Practical Machine Learning and Techniques*. Elsevier.

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

Zareva, A., **P. Schwanenflugel**, and **Y. Nikolova**. 2005. 'Relationship between lexical competence and language proficiency: Variable sensitivity,' *Studies in Second Language Acquisition* 27/4: 567–95.

NOTES ON CONTRIBUTORS

Cynthia Berger is a PhD candidate in the Department of Applied Linguistics & ESL at Georgia State University. Her research interests focus on second language acquisition, corpus linguistics, and psycholinguistics. She is especially interested in how corpus-based and NLP tools can provide insight into lexical acquisition and processing. *Address for correspondence*: Department of Applied Linguistics & ESL, Georgia State University, P.O. Box 4099, Atlanta, Georgia 30302-4099, USA. <*cberger@gsu.edu*>

Scott Crossley is an Associate Professor of Applied Linguistics at Georgia State University. His primary research focus is on NLP and the application of computational tools and machine learning algorithms in language learning, writing, and text comprehensibility. His main interest area is the development and use of NLP tools in assessing writing quality and text difficulty.

Kristopher Kyle is an Assistant Professor in department of Second Language Studies at the University of Hawai'i. His research interests include second language writing and speaking, assessment, and second language acquisition. He is especially interested in applying NLP and corpora to the exploration of these areas.