Berger, C. M., Crossley, S. A., & Kyle K. (2017). Using Novel Word Context Measures to Predict Human Ratings of Lexical Proficiency. *Educational Technology & Society, 20* (2), 201–212.

Using Novel Word Context Measures to Predict Human Ratings of Lexical Proficiency

Cynthia M. Berger[1*], Scott A. Crossley[1] and Kristopher Kyle[2]
[1]Georgia State University, Atlanta, GA, USA // [2]University of Hawai'i at Manoa, Honolulu, HI, USA //
cberger@gsu.edu // scrossley@gsu.edu // kkyle@hawaii.edu
[*]Corresponding author

ABSTRACT

This study introduces a model of lexical proficiency based on novel computational indices related to word context. The indices come from an updated version of the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) and include associative, lexical, and semantic measures of word context. Human ratings of holistic lexical proficiency were obtained for a spoken corpus of 240 transcribed texts produced by second language (L2) adult English learners and native English speakers (NESs). Correlations between lexical proficiency scores from trained human raters and contextual indices were examined and a regression analysis was conducted to investigate the potential for contextual indices to predict proficiency scores. Four indices accounted for approximately 42% of the variance in lexical proficiency scores in the transcribed speech samples. These indices were related to associative, lexical, and semantic operationalizations of word context. The findings demonstrate that computational measures of word context can predict human ratings of lexical proficiency and suggest that lexical, semantic, and associative context each play an important role in the development of lexical proficiency.

Keywords

Second language acquisition, Lexical proficiency, Word context, Natural language processing, Vocabulary

Introduction

While most researchers agree that lexical proficiency is an important component of second language (L2) language competence and academic achievement (Alderson, 2005; Daller, Van Hout, & Treffers-Daller, 2003; Laufer, 1992), the construct of lexical proficiency itself is still poorly understood and the field of L2 research lacks a unified theory of vocabulary acquisition (David, 2008; Schmitt, 2010). This is troubling given the need to understand how L2 lexicons develop in order to allow principled decisions regarding language pedagogy, student placement, and curricula.

Past studies investigating L2 lexical proficiency have examined the intrinsic difficulty of lexical items (e.g., Laufer, 1997), the development of lexical automaticity (e.g., Hulstijn, Van Gelderen, & Schoonen, 2009), receptive vs. productive lexical knowledge (e.g., Melka, 1997), and the distinction between *breadth* and *depth* of knowledge (e.g., Read, 2000). Another approach to conceptualizing and assessing L2 lexical proficiency examines the manner in which L2 lexical items are stored, processed, and retrieved from the mental lexicon (Aitchison, 1994). The assumption behind such an approach is that newly acquired lexical items will need to assimilate into a network of already known words, resulting in restructuring of the network as whole. Lexical proficiency is thus understood as the ability to not only differentiate between semantically related words, but to recognize the variety of ways in which lexical items may be connected to one another (Read, 2004; Singleton, 1999). Presumably, the L2 lexicon strengthens as learners develop stronger links between items and are able to more easily accommodate new words in the network (Haastrup & Henriksen, 2000). The traditional approach to investigating mental lexicons analyzes online language processing in order to gain insights into the mental lexicon and how lexical items are stored, processed, and retrieved (e.g., Conklin & Schmitt, 2008; Ellis & Beaton, 1993; Laufer, 1997). While word frequency is generally considered one of the best predictors of language processing (Balota & Chumbley, 1984; Whaley, 1978), in the current study we investigate the variability of word context to better understand lexical proficiency from a network perspective.

A promising method for better understanding L2 lexical proficiency lies in the use of natural language processing (NLP, Meurers, 2013) tools, such as the Tool for the Automatic Analysis of Lexical Sophistication (Kyle & Crossley, 2015), Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) and AntWordProfiler (Anthony, 2014). Such NLP analytics allow researchers and educators to operationalize many of the constructs related to lexical proficiency and quantify them in learner-produced data. The aim of the current study is to determine if computational indices related to associative, lexical, and semantic word context can increase our understanding of L2 lexical knowledge. To do so, we analyze the spoken lexical output of both L2 English learners and native English speakers in relation to human ratings of lexical proficiency using five novel indices made available in a recently updated version of TAALES (version 2.0).

The key questions that motivate our study are as follows:

- What is the relationship between word context and human ratings of lexical proficiency?
- Can measures of lexical, semantic, and associative word context predict human ratings of lexical proficiency?

Our goal is to predict human ratings of holistic lexical proficiency in spoken language to better understand the construct of lexical proficiency and the role that word context may play in identifying and predicting lexical proficiency in L2 populations.

Background

Natural language processing analytics and lexical proficiency

An emerging approach to better understanding lexical proficiency involves the computational analysis of language produced by language learners (Meurers, 2013). In such an approach, NLP tools are used to analyze large samples of learner-produced text (spoken or written) with the goal of gaining further insight into learner language and the development of lexical proficiency. Linguistic features investigated in learner output typically include psycholinguistic word properties, such as frequency and familiarity (Balota et al., 2004), age of acquisition (Ellis & Morrison, 1998), word associations (Nelson & Friedrich, 1980), and imageability, concreteness, and meaningfulness (Altarriba, Bauer, & Benvenuto, 1999; Paivio, 1991). NLP approaches have demonstrated strong relationships between computational indices and human ratings of lexical proficiency or at predicting the development of lexical proficiency over time. For example, NLP studies of L2 lexis have shown that more lexically proficient L2 learners produce less concrete words, less specific words, a greater variety of words, less frequent words, less familiar words, and words with more senses (Crossley & McNamara, 2013; Crossley, Salsbury, & McNamara, 2009; Crossley, Salsbury, & McNamara, 2013; Crossley, Salsbury, McNamara, & Jarvis, 2010; Crossley, Salsbury, McNamara, & Jarvis, 2011; Kyle & Crossley, 2015).

Still, the use of computational indices to predict lexical proficiency remains incomplete, in part because NLP approaches are often limited to the analysis of individual words produced by learners (e.g., word frequency or concreteness). Such approaches do not give researchers insight into how lexical items are understood in relation to other items in the mental lexicon. One potential solution is to develop computational indices that quantify the relationship of individual words to other words in the lexicon. This approach is taken up by the current study and described further below.

The role of context

Historically, one of the most consistently successful predictors of lexical proficiency and growth over time has been word frequency (Balota & Chumbley, 1984; Whaley, 1978). As learners gain lexical proficiency, they use words that are less frequent in everyday usage (Ellis, 2002a; Ellis, 2002b; Sorrell, 2013). The frequency data upon which such research is based are typically word occurrences derived from large reference corpora. However, this manner of operationalizing frequency has been criticized for treating language as a randomly ordered collection of words, an approach that assumes complete independence of items in the lexicon (McDonald & Shillcock, 2001). Some researchers (e.g., Adelman et al., 2006; Brysbaert & New, 2009a) have suggested that the diversity of contexts in which a word is encountered and the subsequent constraints that context puts on a word's meaning and use may offer more psychologically valid explanations of the word frequency effect.

Anderson's (1991) *principle of likely need* presumes that the goal of human memory is to be efficient while fulfilling goals within a specific environment. Thus, items stored in memory are not equally necessary in any given context. Rather, each item has a need-probability based on past use and current context, with those items that are most likely to be needed more easily available. From this argument, if we assume that the environment in which a word is encountered is a *context*, we can infer that words more likely to be needed in a variety of contexts will be more readily available. This context-oriented perspective on frequency-based learning predicts that the context(s) in which an item is encountered will influence its subsequent retrieval and production. However, as previously discussed, a word's context is not captured by the frequency measures commonly used in investigations into language acquisition. It may be that a more context-oriented construct, implicitly predicted by Anderson's the *principle of likely need*, has explanatory power in this arena as well.

While the current study seeks to explore this possibility, we are not the first to do so. One approach to acknowledging the role of context in word representation has been to define context broadly as the number of documents or genres in which a lexical item is typically encountered. For example, Adelman et al. (2006) used the term *contextual diversity* to describe the variability of contexts (i.e., documents) in which a word occurred across three different corpora. The authors demonstrated that contextual diversity predicted word processing times in NES online psycholinguistic tasks independent of frequency and regardless of variables related to concreteness, imageability, and ambiguity. Adelman et al.'s findings were substantiated by Brysbaert and New (2009b), who offered a similar measure derived from a corpus of film and television subtitles. Brysbaert and New demonstrated that their measure (i.e., the number of films or television shows in which a word occurred) was more predictive of NES lexical decision response times than a word frequency measure alone.

The majority of work on contextual diversity has explored how L1 subjects retrieve and process words, rather than investigating L2 acquisition or lexical proficiency. In an exception, Kyle and Crossley (2015) demonstrated that approximately 26% of the variance in spoken lexical proficiency ratings could be explained by a contextual diversity index derived from the written BNC. The correlation between contextual diversity and lexical proficiency scores was negative in their study, suggesting that raters' impressions were positively impacted by speakers who used words occurring in fewer contexts. Another approach to word context and language proficiency was taken by Crossley, Subtirelu, and Salsbury (2013), who used word association norms to investigate about 100 nouns and verbs most frequently produced by beginning-level English learners. While the authors did not find that this operationalization of context predicted the words learners produced, their basic approach is one that we adopt and expand in the current study.

Current study

In the majority of contextual diversity studies described above, context is defined as either an individual document or an entire genre. In this regard, a large-grained approach is used to conceptualize context, one that ignores a word's lexical and semantic environments, as well as its affiliates in the mental lexicon. In the current study, we adopt the term *contextual distinctiveness* (McDonald & Shillcock, 2001) to refer to the constraints put on a word by its immediate lexical and semantic context. We also employ the term to address context in cognitive representation. While *contextual distinctiveness* was originally coined to refer to a word's proximate lexical environment, its use in this paper is intended to more broadly reference not only lexical context, but associative and semantic word contexts as well. We do this by examining the degree to which computational indices quantifying the role of word context in language usage and cognitive representation may predict human ratings of lexical proficiency. We predict that these measures may serve to quantify the unique role that contextual distinctiveness plays in defining L2 lexical acquisition and proficiency.

We analyzed a corpus of transcribed speech samples using indices made available in the text analysis tool TAALES 2.0 (Kyle & Crossley, 2015). In order to capture a wide variety of lexical proficiency, we analyzed speech samples from 180 L2 learners at three different levels of proficiency and an additional 60 speech samples produced by native English speakers (NESs). Trained raters scored the speech samples using a holistic lexical proficiency rubric. Prior to statistical analysis, we divided the scored speech samples into a training and a test set. We then conducted correlational and linear regression analysis on the training set to examine the relations between the human lexical proficiency scores and the TAALES indices. The same model was then extended to the test set in order to cross-validate the model and ensure generalizability.

Method

Corpus

The corpus analyzed in this study contained transcribed, spoken data collected from both L2 and NESs (Crossley et al., 2010). In the L2 sub-corpus, transcribed data were derived from naturalistic, interactional discourse between an English learner and a NES interlocutor. L2 learners were either matriculated undergraduates or intensive English program students at two different universities in the United States. First language (L1) backgrounds of the learners included Arabic, French, Turkish, Japanese, Korean, Mandarin, and Spanish. Prior to participating in the study, learners were grouped into beginning (n = 60), intermediate (n = 60), and advanced (n = 60) proficiency levels based on the Test of English as a Foreign Language (TOEFL) or ACT ESL Compass scores, for a total of 180 L2 speech samples collected across levels. Proficiency levels were used to ensure a range of proficiency variance in the speech samples that comprise the corpus; however, the levels themselves

were not used as dependent variables in the current study. The NES corpus was comprised of 60 speech samples selected from the *Switchboard* corpus (Godfrey & Holliman, 1993; see Crossley et al., 2010 for details), a collection of approximately 2,400 telephone conversations taken from naturalistic conversations between 543 NESs across the United States. In total, the corpus analyzed in the current study included 240 speech samples.

Human ratings of lexical proficiency

Human ratings were based on transcribed interactions between a given speaker and his or her interlocutor. The corpus was divided into segments of the speech that contained approximately 150 words each for the speaker of interest and captured a complete interaction. Three trained raters assessed the 240 speech samples for lexical proficiency using a holistic grading rubric based on a 5-point Likert-scale, with a score of 5 demonstrating skillful, consistent mastery of the English lexicon and a score of 1 indicating little lexical mastery. The rubric was developed based on an adaptation of holistic proficiency rubrics produced by American College Testing (ACT), the College Board, and the American Council on the Teaching of Foreign Languages' (Breiner-Sanders, Lowe, Miles, & Swender, 2000). In order to evaluate inter-rater reliability, Pearson correlations between all possible pairs of raters' responses were averaged and weighted. For the full corpus, the average correlation among raters was $r = .808$ ($p < .001$) with a weighted correlation of $r = .927$. See Crossley et al. (2010) for details regarding rubric development and rater calibration.

Word context measures

Five measures that take into account the role of context in word representation were included in the current study as potential predictors of L2 lexical proficiency. These are listed in Table 1 and explained further below. All the indices used were obtained using TAALES 2.0 (for details, see Kyle & Crossley, 2015). The five indices were selected to operationalize the construct of contextual distinctiveness according to associative, lexical, and semantic principles.

Table 1. TAALES 2.0 indices measuring word content

Index	Abbreviation	Description	Subconstruct	Reference
Edinburgh Associative Thesaurus response type and token counts	EAT_types	Number of word types elicited by target word	Associative context	Kiss et al. (1973)
	EAT_tokens	Number of work tokens elicited by target word		
University of South Florida stimuli counts	USF	Number of stimuli words that resulted in production of word	Associative context	Nelson et al., (1998)
Semantic ambiguity	SemD	Variability of contexts in which word occurs	Semantic context	Hoffman et al. (2013)
Relative entropy	McD	Amount of information conveyed by word about its frequent lexical contexts	Lexical context	McDonald & Shillcock (2001)

Associative context: Word association indices

One approach to quantifying the distinctiveness of a word's context is to calculate the number of other words commonly associated with it. The motivation behind incorporating word association (WA) measures in the current study was the assumption that words with a greater number of associations would be less contextually distinct. TAALES 2.0 indices derived from two publicly available behavioral datasets were used to operationalize associative context. These are described below.

Edinburgh Association Thesaurus response type and token counts

The Edinburgh Associative Thesaurus (EAT) index calculates the number of words produced in response to target stimuli in a written WA task. Norms are derived from NES subjects' associative responses to 8400 English words (Kiss, Armstrong, Milroy, & Piper, 1973). Because the EAT index measures the number of associates a

word elicits, a higher number indicates a wider variety of associates. In the current study, it was assumed that words with a greater number of response associations are less contextually distinct. For example, the word *worry,* which elicited 65 different associate types in EAT data, may be less contextually distinct than the word *husband,* which elicited only 15. TAALES 2.0 reports both type and token counts for EAT responses. Type counts reflect the number of unique words elicited in response to a given stimulus, while token counts reflect the number of responses elicited in total.

University of South Florida stimuli counts

The University of South Florida (USF) association norms (Nelson, McEvoy, & Schreiber, 1998) calculate WAs in a reverse manner from the EAT index by reporting the number of stimuli words that resulted in production of the target word as an associate in a WA task. Association data was collected from NES subjects in response to 5,019 stimulus words, resulting in a total of 10,470 response words. Words that rank high on this measure are considered more accessible and are thus more likely to come to mind in response to a variety of cues (see Nelson et al., 1998 for details). For example, the word *love* was produced in response to 181 different stimuli and may be relatively less contextually distinct than a word like *bride*, which was produced in response to only 6 stimuli in the USF WA task.

Lexical context: Relative entropy

The TAALES 2.0 relative entropy index (McD) used in the current study is derived from values calculated by McDonald and Shillcock (2001) for 8,000 English lexemes in the spoken BNC (2007). McD is based on the probability of a word co-occurring with other words in general language usage. The McD index reports the amount of information conveyed by a word about its frequent lexical contexts. A word with a higher McD value, such as *lone* (3.748), occurs in more distinct lexical contexts, while a word with a lower value, such as *today* (0.16), occurs in a variety of lexical contexts (i.e., is likely to co-occur with a variety of other words). In this example, the word *lone* is considered more informative about its contexts of lexical occurrence than *today*. In the current study, we assume that words with higher McD values are also more lexically contextually distinct.

Semantic context: Semantic ambiguity

The TAALES 2.0 semantic diversity (SemD) index operationalizes a word's semantic ambiguity based on the variability of semantic contexts in which it occurs. This computational measure is based on Latent Semantic Analysis (Landauer et al., 1998; LSA) and was originally calculated by Hoffman, Ralph, and Rogers (2013) based on analysis of 1,000-word "contexts" in the written BNC (2007). The SemD index includes values for 31,739 English words. A high SemD value for a word is assumed to be more contextually variable, or ambiguous (i.e., occurring in a variety of semantic contexts), than a lower SemD value. For example, the word *time* has a SemD value of 2.30 and is thus more semantically ambiguous than the word *puppy*, which only has a SemD value of 0.93. The assumption motivating inclusion of SemD in the current study is that words with lower SemD values are more semantically contextually distinct than words with higher SemD values.

Analysis

Prior to conducting a stepwise linear regression, indices were checked for normal distribution. Correlations were then conducted to examine the relations between the proposed context indices and human ratings of lexical proficiency. Only indices that demonstrated a correlation of at least $r > .100$ were retained. If two or more indices demonstrated strong multicollinearity ($r > .900$), only the index that correlated the strongest with lexical proficiency would be retained. In order to cross-validate our model and assess its generalizability, we divided spoken texts into training and test sets. The training set comprised approximately 67% ($n = 166$) of the texts while the test set contained approximately 33% ($n = 74$) of the texts (Witten, Frank, & Hall, 2011). If a model derived from a training set predicts the dependent variable in a test set at a similar accuracy rate, the model can be considered stable and generalizable to the population.

Results

Correlations and normality checks

None of the indices deviated from normal distribution. All five of the indices demonstrated a correlation of at least $r > .100$ with lexical proficiency ratings (Table 2). No indices demonstrated multicollinearity with other indices. Thus, all five the indices were retained for regression analysis.

Regression analysis

Training set

A stepwise linear regression using the five indices yielded a significant model, $F(4,161) = 28.713$, $p < .001$, $r = .645$, $r^2 = .416$. Four indices were significant predictors in the regression: USF, SemD, McD, and EAT_types. Two of these indices had been selected to operationalize associative context (USF and EAT_types), while SemD captures semantic context and McD indexes lexical context. EAT_tokens was not a significant predictor and was not included in the model.

Table 2. Correlations between lexical proficiency scores and word context indices

Index	*r*	*p*
USF	-.534	0
SemD	.475	0
McD	.318	0
EAT_tokens	.191	.003
EAT_types	.116	.073

The results of the regression model (Table 3) demonstrate that the combination of four contextual indices accounts for roughly 42% of the variance in human judgments of lexical proficiency for the 166 essays that comprised the training set. The standardized coefficients (β) in Table 3 indicate the number of standard deviation (SD) changes we would expect in lexical proficiency for a one SD change in any given index. For example, for every one SD increase in USF value for a given speech sample, we can expect to see a 0.334 SD decrease in lexical proficiency ratings.

Table 3. Linear regression results for contextual indices

Entry	Index	*r*	R^2	R^2 change	β	*SE*	B	*T*
1	USF	.502	.252	.252	-.334	0.009	-0.046	-4.944
2	Sem_D	.591	.349	.097	0.350	1.346	7.207	5.356
3	McD_CD	.630	.397	.048	0.216	0.550	1.885	3.430
4	Eat_Types	.645	.416	.019	0.140	0.032	0.073	2.297

Note. β = standardized; B = unstandardized β; Estimated constant term is -14.732; all *t* significant at < .05.

Test set

The model for the test set yielded $r = .702$, $r^2 = .493$. These results indicate that the four indices retained in the training set accounted for roughly 49% of the variance in human ratings of lexical proficiency for the 74 texts that comprised the test set.

Discussion

The purpose of this study was to investigate links between novel computational word context indices and human ratings of lexical proficiency. The results indicate that there are strong relationships between contextual distinctiveness and lexical proficiency. The findings have important implications for second language acquisition (SLA) and lexical proficiency because they suggest that lexical, semantic, and associative context may each contribute to the development of lexical proficiency.

Our first research question asked whether there was a relationship between word context and human ratings of lexical proficiency. The findings suggest that there is indeed such a relationship, with all contextual

distinctiveness indices showing significant correlations with lexical proficiency (Table 2). USF demonstrated a large effect size ($r > .50$) in its correlation with lexical proficiency. In addition, two other indices (SemD and McD) demonstrated medium effect sizes ($r > .30$), while two (EAT_tokens and EAT_types) demonstrated small effect sizes ($r > .10$) (Cohen, 1988).

Interestingly, an inverse relationship between semantic ambiguity (SemD) and lexical variability (McD) was reported: As speakers become more lexically proficient, they use words that are more semantically ambiguous but also more lexically distinct. Thus, more proficient speakers use words that occur in a wide variety of semantic contexts (higher SemD) but are less likely to co-occur with a wide variety of words (higher McD). This finding suggests that semantic and lexical context are distinct subconstructs that may capture unique aspects of contextual distinctiveness. There was also an inverse relationship between EAT response type and token counts and USF stimuli counts. Recall that both sets of WA measures were selected to operationalize associative word context (i.e., how words are related to one another in the mental lexicon): The former reports the number of response to a target word while the latter reports stimuli counts. Because the relationship between USF and lexical proficiency was significantly stronger than the relationship between EAT and lexical proficiency (Tables 2 and 3), it could be that stimuli counts (e.g., USF norms) represent a more robust manner of quantifying associative context than the analysis of response counts. The existence of inverse relationships among indices suggests that each measures distinct subconstructs of contextual distinctiveness. This explanation is elaborated below when we analyze speech samples with reference to the four indices.

Our second research question asked whether contextual distinctiveness measures could predict human ratings of holistic lexical proficiency. Indeed, a combination of four indices related to word context (Table 3) explained nearly 42% of the variance in lexical proficiency in transcribed speech samples. This model retained indices pertaining to each operationalization of word context: associative, lexical, and semantic. Relationships between lexical proficiency and the indices selected by our model are outlined in Table 4 and further explained below.

Table 4. Relationship between indices retained by model and lexical proficiency

Measure	Subconstruct	Variance explained	More proficient speakers…
USF	Associative context	25%	Used words associated with a smaller number of stimuli in WA tasks
SemD	Semantic context	10%	Used words that occur in a wider variety of semantic contexts
McD	Lexical context	5%	Used words that occur in more distinct lexical contexts
EAT_types	Associative context	2%	Used words that produce a variety of responses in WA tasks

The greatest predictor of lexical proficiency in speech samples was USF stimuli counts, which accounted for over 25% of the variance in human ratings. The correlation between USF and lexical proficiency ratings was negative, suggesting that as speakers gain lexical proficiency they use words that are less likely to be produced in response to a range of stimuli in free association tasks. The second strongest predictor of lexical proficiency was SemD, which explained another 10% of variance in ratings. SemD was positively correlated with lexical proficiency, meaning that more lexically proficient speakers produce words that occur in a wider variety of semantic contexts. The third strongest predictor of lexical proficiency, McD, explained roughly 5% of lexical proficiency scores. McD reports the amount of information conveyed about a word's frequent lexical contexts in language usage, with a higher value indexing words that statistically co-occur in the company of fewer words. Because this measure positively correlated with ratings, we know that more proficient speakers are more likely to produce words that occur in more distinct lexical contexts. The final predictor in our model, EAT_types, explained 2% of the variance in holistic lexical proficiency ratings. This index was positively correlated with lexical proficiency and suggests that more proficient speakers produce words that elicit a greater variety of responses in word association (WA) tasks.

To demonstrate how the above model reflects lexical proficiency in actual language use, we offer a comparison of two contrasting speech samples from the corpus. Sample A received the highest rating possible from human raters, while Sample B received one of the lowest. Certain words have been italicized to highlight their reference in discussion below.

Sample A

Okay. I don't really, I more, I don't know about the government as much as the people. I wouldn't consider to be a threat at all and I really don't feel much like the Soviet Union itself is a threat anymore. I'm, I'm worried about them. They're in a very tumultuous state right now with the kinds of adaptations that they're attempting to go through. Yeah. Yeah I think that's… that's a real important *aspect* and that as the… the… the most the mo-… let's see, the more that we do that we do or that we can do to help them become self-sufficient is going to *eliminate* more of the risk of that becoming, you know, a *reality*. I know that this last winter was very hard on several areas in the… in the… the Ukraine, particularly the coal mining regions of Siberia. The people there have money that's not their problem, but there's no food for them to buy and it's, you can't eat money.

Sample B

Okay. I think *one* year. No here study English I think when? Ah, no. I am high *school*. I *class*… eh… English… eh… in my no study much English no… no luck. Ah, I study English now. It is very important for my work. I am account… account? Accountant. Okay, in Caracas. I am tax *manager*. Is very, very important English. I am customer Proctor and Gamble. General Motors. Chrysler motors. Americans… ah… they speak English and me speak English. No, no. They… ah… to go in Venezuela, no learn English. No problem they no English. Ours? Okay. Do you think they should speak Spanish? Do you think they should speak Spanish? Speak Spanish? No, they no.

Table 5 reports the speech samples' holistic proficiency ratings, as well as their scores for each of the four indices retained by our model. Index values are reported in z-scores, which adjust a given distribution to have a mean of 0 and a standard deviation of 1. Figure 1 represents these z-scores visually.

Table 5. Z-scores for Samples A and B context indices

Speech sample	Lexical proficiency rating	Z-scores			
		USF	Sem_D	McD_CD	EAT_types
A	5.000	-0.545	1.596	1.368	1.438
B	1.667	0.715	-2.115	0.242	-0.329

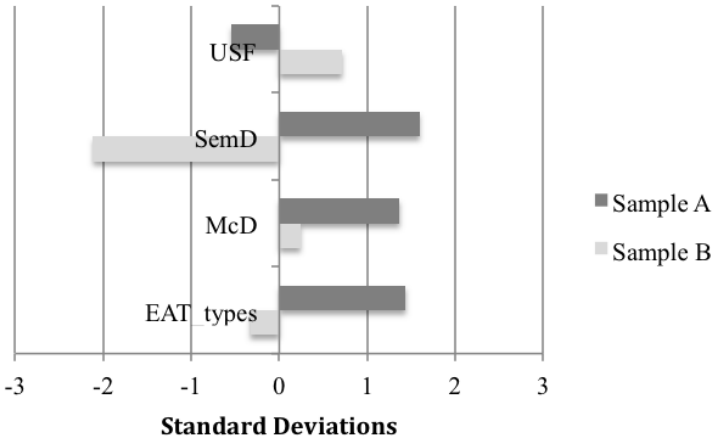

Figure 1. Z-scores for speech Samples A and B

The two speech samples analyzed here performed as predicted by our model, with the exception of McD_CD in Sample B, which was slightly above the group mean (rather than below, as predicted). However, it was still lower than Sample A's McD_CD score. Because USF correlated negatively with lexical proficiency in our model, the highly proficient speaker that produced Sample A used fewer words elicited by a range of stimulus words in the USF WA task. For instance, Sample A contains words such as *aspect* (1 USF) and *reality* (6 USF), while sample B used more words elicited by a range of different stimulus words, such as *school* (183 USF) and *class* (95 USF). Unlike USF, EAT types demonstrated a positive relationship with lexical proficiency. For this reason, the more highly rated Sample A contained more words that elicited a greater number of response types in the EAT WA task. This finding runs counter to the negative relationship between USF and lexical proficiency. Most likely, this is because USF and EAT measure different aspects of lexical association: EAT reports the number of responses to a target word while USF reports the number of stimuli that elicited the word. For example, *reality*, used in Sample A, has a low USF value (6) but a moderately high EAT_types value (43). In

other words, *reality* is unlikely to be elicited by a variety of stimuli in WA tasks, even though it elicits 43 different response types when used as a stimulus. Sample B, on the other hand, contains words like *one*, which has a high USF value (63) but a relatively low EAT_types value (23). Meanwhile, other words used in Sample B, such as *school*, have moderate EAT_types values (56) but extremely high USF values (183). This is because a word like *school* is elicited by a number of different stimuli in WA tasks but only produces a moderate number of responses when acting as a stimulus.

Table 5 and Figure 1 also indicate that Sample A had higher SemD and McD values than Sample B, as predicted by our model, suggesting that more proficient speakers produce words that are more semantically diverse at the same time that they are more lexically distinct. For example, Sample A contains *eliminate*, which has both a high SemD value (2.09) and a high McD value (1.787). While *eliminate* is likely to occur in a variety of different semantic contexts, it is more constrained in terms of words that it is likely to co-occur with in general language usage. Sample B, however, contains words like *manager*, which occurs in more distinct semantic contexts (SemD = 1.68) but is less lexically distinct (McD = .604).

The associative, lexical, and semantic context indices in our model each explained unique variance in the ratings of lexical proficiency. In addition, none of the indices were highly inter-correlated (the highest correlation was between USF and SemD: $r = -.337$, $p > .001$). These results demonstrate that the measures introduced in our study capture distinct aspects of word context and offer initial evidence to support their independence as subconstructs of word context. This finding also indicates that the development of lexical networks may be impacted independently by the semantic contexts of newly acquired words, the associative relatedness between known and newly acquired words, and the number of statistically frequent lexical neighbors that occur in the company of a given word in everyday language use.

Our results support Kyle and Crossley's (2015) findings, who demonstrated that a contextual diversity measure (the number of individual documents in which a word was found in a reference corpus) predicted roughly one-quarter of the variance in holistic ratings of spoken lexical proficiency. However, the approach used in Kyle and Crossley examined context globally (across texts) and not locally (within a small window of words). Only one other study, to our knowledge, has taken this approach (Crossley, Subtirelu, & Salsbury, 2013), and our findings run contrary to theirs (i.e., USF norms were not predictive of the words produced by beginning-level English learners). This likely results from Crossley, Subtirelu, and Salsbury (2013) only analyzing the most frequently produced nouns and verbs in their learner corpus.

Our findings have the potential to offer insights into L2 testing and assessment, including the automatic scoring of language skills. They indicate that word properties beyond word frequency account for human judgments of lexical proficiency. The findings also suggest that L2 instructional approaches may benefit from methods that move beyond the study of words in isolation to include word context, with emphasis on associative relationships among words and the frequent semantic and lexical contexts of language in use. For instance, one pedagogical activity that our results support is semantic word mapping, which encourages learners to brainstorm and diagram associations between words (Johnson & Steele, 1996; Laufer, 1990). Another approach to incorporating context in language teaching would involve the use of corpus-based concordance tools, which allow learners to observe target words in their most common lexical and semantic environments (Reppen, 2010; Römer, 2008).

Conclusion

The results of this study suggest that computational indices quantifying word context can be used to predict human ratings of spoken lexical proficiency. Our analysis of a corpus of transcribed speech samples included five indices made available in TAALES 2.0. The indices were selected to reflect the construct of word context according to associative, lexical, and semantic operationalizations. Four of these indices demonstrated a significant relationship to holistic ratings of spoken lexical proficiency. A model obtained in stepwise linear regression explained 42% of the variance in human ratings, with indices pertaining to each operationalization of lexical proficiency contributing to the model. Results suggest that computational measures of lexical, semantic, and associative context each play an important role in understanding lexical proficiency.

While our model has predictive validity, the indices themselves cannot be interpreted as contributing to lexical proficiency per se. Lexical proficiency itself is undoubtedly explained by a number of factors. For example, variables related to rhetorical organization, pragmatic and content knowledge, and accuracy undoubtedly impact impressions of lexical proficiency, and these are not included in the current research. Follow-up studies are needed in order to determine the causal role the proposed indices may play in explaining lexical proficiency and

the degree to which these context-related indices relate to other variables known to impact lexical proficiency. Additionally, while we have made claims about the development of proficiency, the current study is based on a cross-sectional dataset. The use of these same computationally derived context measures on a longitudinal dataset would offer greater insights into the relationship between contextual distinctiveness and the development of lexical proficiency in learners over time (Ortega & Iberri-Shea, 2005).

Nonetheless, the novel computational indices proposed in the current study offer a unique manner with which to investigate the role of contextual distinctiveness in the mental lexicon (associative) and in language use (semantic and lexical). As tools for the enhanced application of learner analytics, they also may allow researchers and educators to better analyze learner-produced data in order to assess learner progress and predict future performance.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determined word-naming and lexical decision times. *Psychological science, 17*, 814-823.

Aitchison, J. (1994). *Words in the mind: An Introduction to the mental lexicon*. Oxford, UK: Blackwell.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The Interface between learning and assessment*. London, UK: Continuum.

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers, 31*, 578-602.

Anderson, J. R. (1991). The Place of cognitive architectures in a rational. *Architectures for Intelligence*, 1.

Anthony, L. (2014). *AntWordProfiler* (Version 1.4.1) [Computer software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The Role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*(3), 340-357.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*(2), 283-316.

BNC. (2007). The British Naitonal Corpus (version 3, BNC XML Edition). Retrieved from http://www.natcorp.ox.ac.uk/

Breiner-Sanders, K. E., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines—Speaking: Revised 1999. *Foreign Language Annals, 33*(1), 13-18.

Brysbaert, M., & New, B. (2009a). 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), 977-990.

Brysbaert, M., & New, B. (2009b). Subtlexus: American word frequencies. Retrieved from http://subtlexus.lexique.org

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72-89.

Crossley, S. A., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307-334.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*(4), 561-580.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45*(1), 182-193.

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. Language *Learning & Technology, 17*(2), 171-192.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M. Daller (Eds.), *Vocabulary Knowledge: Human ratings and automated measures* (Vol. 47, pp. 105-135). Amsterdam, The Netherlands: John Benjamins.

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning. *Studies in Second Language Acquisition, 35*(4), 727-755.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197-222.

David, A. (2008). A Developmental perspective on productive lexical knowledge in L2 oral interlanguage. *Journal of French Language Studies, 18*(03), 315-331.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 515-523.

Ellis, N. C. (2002a). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(02), 143-188.

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Second Language Acquisition, 24*, 297-339.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*(4), 559-617.

Godfrey, J. J., & Holliman, E. C. (1993). *Switchboard-1 [CD1-ROM]*.

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

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics, 10*(2), 221-240.

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*(3), 718-730.

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics, 30*(04), 555-582.

Johnson, D. D., & Steele, V. (1996). So many words, so little time: Helping college ESL learners acquire vocabulary-building strategies. *Journal of Adolescent & Adult Literacy, 39,* 348-357.

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

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Laufer, B. (1990). Ease and difficulty in vocabulary learning: Some teaching implications. *Foreign Language Annals, 23,* 147-155.

Laufer, B. (1992). Reading in a foreign langauge: 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 & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 140-155). Cambridge, MA: Cambridge University Press.

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*(3), 295-323.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 84-102). Cambridge, MA: Cambridge University Press.

Meurers, D. (2013). Natural language processing and language learning. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1-13). New York, NY: Blackwell Publishing Ltd.

Nelson, D. L., & Friedrich, M. A. (1980). Encoding and cuing sounds and senses. *Journal of Experimental Psychology: Human Learning and Memory, 6*(6), 717-731.

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*(3), 402-407..

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics, 25*, 26-45.

Paivio, A. (1991). *Images in mind: The Evolution of a theory*. New York, NY: Harvester Wheatsheaf.

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

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209-228). Amsterdam, The Netherlands: John Benjamins Publishing.

Reppen, R. (2010). *Using corpora in the language classroom*. New York, NY: Cambridge University Press.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 112-130). Berlin, Germany: Mouton de Gruyter.

Schmitt, N. (2010). *Researching vocabulary: A Vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan.

Singleton, D. M. (1999). *Exploring the second language mental lexicon*. Cambridge, MA: Cambridge University Press.

Sorrell, C. J. (2013). Zipf's law and vocabulary. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Hoboken, NJ: Wiley-Blackwell.

Whaley, C. (1978). Word—nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*(2), 143-154.

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