Available online at www.sciencedirect.com

ScienceDirect

System 41 (2013) 965–981

SYSTEM

www.elsevier.com/locate/system

Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications

Scott A. Crossley [a,*], Tom Cobb [b], Danielle S. McNamara [c]

[a] *Department of Applied Linguistics/ESL, Georgia State University, 34 Peachtree St. Suite 1200, One Park Tower Building, Atlanta, GA 30303, USA*

[b] *Département de didactique des langues, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal QC H3C 3P8, Canada*

[c] *Department of Psychology, Arizona State University, P.O. Box 872111, Tempe, AZ 85287-2111, USA*

Received 10 October 2012; revised 30 July 2013; accepted 9 August 2013
Available online

Abstract

In assessments of second language (L2) writing, quality of lexis typically claims more variance than other factors, and the most readily operationalized measure of lexical quality is word frequency. This study compares two methods of automatically assessing word frequency in learner productions. The first method, a band-based method, involves lexical frequency profiling, a procedure that first groups individual words into families and then sorts these into corpus-based frequency bands. The second method, a count-based method, assigns a normalized corpus frequency count to each individual word form used, yielding an average count for a text. Both band and count-based methods were used to analyze 100 L2 learner and 30 native speaker freewrites that had been classified according to proficiency level (i.e., native speakers and beginning, intermediate and advanced L2 learners). Machine learning algorithms were used to classify the texts into their respective proficiency levels with results indicating that count-based word frequency indices accurately classified 58% of the texts while band-based indices reported accuracies that were between 10% and 22% lower than count-based indices.

Keywords: Frequency analysis; Frequency lists; Band-based frequency measures; Count-based frequency measures; Computational linguistics; Lexical sophistication; Active and passive lexical proficiency; Learner corpora

1. Introduction

Computational text analysis has produced a number of indices that have proven useful in language learning contexts. These range from simple measures of sentence or *t*-unit length (Hunt, 1965), to complex measures of cohesion, grammatical development, and lexical sophistication (Crossley et al., 2010, 2011a, 2011b; Lu, 2011; McCarthy and Jarvis, 2010). Such measures have been used to assess text readability (Crossley et al., 2008), grade

* Corresponding author.
E-mail addresses: sacrossley@gmail.com, scrossley@gsu.edu (S.A. Crossley), cobb.tom@uqam.ca (T. Cobb), dsmcnamara1@gmail.com (D.S. McNamara).

0346-251X/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.system.2013.08.002

learning materials (Cobb, 2007; Crossley et al., 2007), assess productive lexical proficiency (Laufer and Nation, 1995), and score semi- or fully automatically written and spoken productions in both first (L1) and second language (L2) contexts (Crossley and McNamara, 2012; Grant and Ginther, 2000; Jarvis et al., 2003). Many such implementations have taken place in real world, medium-stakes situations such as student placement or formative assessment. The present study involves the use of computational frequency-based indices to predict the proficiency levels of L1 and L2 writers. The study focuses particularly on the types of frequency indices available and their practicality for analyzing learning production. Specifically, we examine the relative accuracy of two frequency approaches (band-based and count-based frequency approaches) to predict the proficiency level of written samples produced by L1 writers of English and L2 writers at beginning, intermediate, and advanced proficiency levels.

1.1. Literature review

Large corpora, accessible in recent years, have rendered frequency ratings of word-forms more readily available. For example, a word such as *the* can be quantified with relative ease as a highly frequent word with 6,041,234 instances in the 100-million word British National Corpus (BNC) compared to *analysis* with 13,118 instances. Large digital corpora make it possible to calculate the frequency value of each text word in a corpus and assign an average rating to the whole text, so that a single number or small set of numbers can indicate its lexical sophistication.[1] A more sophisticated text by this reasoning is one with more low-frequency words.

There is empirical and theoretical support for frequency as a reasonably reliable and valid operational stand-in for lexical knowledge. From a receptive perspective, studies have supported the notion that high frequency words are recognized (Kirsner, 1994) and named more rapidly (Balota and Chumbley, 1984; Forster and Chambers, 1973). From a production perspective, Ovtcharov et al. (2006) found a significant difference for vocabulary frequency in passing and failing transcribed oral interviews in a high stakes civil service language test where "well developed vocabulary" was one of the qualitative criteria for success. Crossley and Salsbury (2010) demonstrated that the frequency of a word is an important element in predicting whether beginning level L2 learners will produce that word, with the understanding that less complex words are produced first. Under the same premise, studies have shown that lower-level L2 learners produce words of higher frequency in writing (Bell, 2003; Laufer and Nation, 1995; Crossley et al. 2011a) and speaking (Crossley et al. 2011b) than higher-level L2 learners. The production of more frequent words in writing is also predictive of writing proficiency with essays scored as low proficiency containing more frequent words than essays scored as high proficiency (Morris and Cobb, 2004, Crossley and McNamara, 2012; Laufer and Nation, 1995).

However, there are also problems with relying on frequency indices as reliable proxies for lexical knowledge. One problem is that frequency indices measure lexical knowledge at the surface code level (i.e., at the text level) as compared to indices that measure knowledge at semantic textbase level (i.e., indices that examine explicit connections and referential links in text) and the situational level (i.e., indices that examine text causality, temporality, inferencing, and given/new information; Crossley and McNamara, 2012; Graesser et al., 1997). In addition, more frequent words have a tendency to be more polysemous as a result of the *law of least effort* which states that language learners, whether of their first or second language, economize vocabulary by extending the number of senses a word contains to conserve lexical storage (Murphy, 2004). Over time, this law leads to the most frequent words containing the most senses (Zipf, 1945) and, as a result, exhibiting greater degrees of ambiguity potentially leading to more processing difficulty (Davies and Widdowson, 1974).

In terms of execution, there are a number of ways that frequency can be calculated and there are different frequency objectives that can be targeted. Major procedural questions involve whether to group lexical units into lemmas or families or neither, and whether to calculate frequency using band-based indices or count-based indices. We define band-based indices as those that calculate word frequency as a function of frequency bands (the bands applicable to adult learners contain words that occur at intervals of a thousand). Examples of these indices include Lexical Frequency Profiles (LFP: Laufer and Nation, 1995) and P_Lex (Meara and Bell, 2001). We define count-based frequency indices as those that calculate word frequency as a function of word incidences as found in large-scale corpora. The best example of this are the CELEX frequency norms (Baayen et al., 1995) reported by Coh-Metrix (Graesser et al. 2004, McNamara and Graesser, 2011). Once a frequency approach has been selected

[1] Word frequency has often been used as a measure of lexical richness (Laufer and Nation, 1995). However, word frequency is only one aspect of lexical richness, which, by definition, includes lexical diversity, lexical sophistication (i.e., frequency), and lexical density (Jarvis, 2012).

(either band- or count-based), questions arise as to whether the goals of the analysis should target accuracy or usability, and productive or receptive language ability.

Probably the best-known band-based frequency measure is the Lexical Frequency Profiles (LFP), which was developed by Laufer and Nation (1995), refined in Nation (2006), and employed extensively in the language-learning world thereafter through websites like Lextutor (www.lextutor.ca). In LFP, frequency is calculated for whole groups of words in two senses. Word families (inflections and obvious derivations) are given a frequency rating based on the sum of the BNC frequencies of all the members. For example, *go* (87,021 occurrences), *goes* (14,264), *going* (63,433), *gone* (18,433), and *went* (45,538) sum to a family frequency of 228,689, putting it well within the most frequent 1000 families (with a frequency greater than 12,639, the cutoff between first and second 1000 lists in recent versions of LFP). This process produces a basic frequency list for a pedagogically practical number of families (for instance, Nation, 2006, targeted 14 thousand families, and Cobb and Horst, 2011 followed Nation's methodology to add 6 thousand more, bringing the total to 20 thousand families on the Lextutor version of LFP). These families are further refined by considerations of range (the first 10 thousand families were represented in all ten of the corpus' 10 million-word subdivisions) and genre (the first 2 thousand families are drawn exclusively from the spoken subdivision). Finally, groupings of 1000 such families are assembled into bands, such that texts can be described and even color-coded by a computer program as comprising, for example, 70 percent first-1000 band word families (or k-1 families), 10 percent k-2 families, and the remainder k-3 to k-7 families, in a typical profile of a newspaper text.

An advantage of this approach is the sense of clarity it can offer to teachers, course designers, novice researchers, and other practitioners. The disadvantages are the potential information loss that comes with grouping, and hence fewer distinctions, as well as the bias toward receptive knowledge that is inherent in the construct and construction of word families, particularly the idea of an "obvious derivation" (a learner may recognize that *electricity* is a member of *electric* without being able to produce this formation). Meara (2005) has also criticized the use of frequency bands to measure production on the grounds these bands cannot be sensitive to small linguistic differences and changes such as those found in developing lexicons.

Nonetheless, the LFP approach has been somewhat successful as a production measure. LFPs successfully distinguish learner productions by proficiency level (Laufer and Nation, 1995), distinguish between success and failure on productive language tests in both speech (Ovtcharov et al. 2006) and writing (Morris and Cobb, 2004), reliably characterize lexical profiles of different text types and predict L2 readers' comprehension of text (Laufer, 1992; Nation, 2006), and serve as the raw material for band-based assessments, both receptive (Beglar and Nation, 2007; Nation, 1990) and productive (Laufer and Nation, 1999), which reliably predict broader language competence (Cobb, 2000; Cobb and Horst, 1994; Meara and Buxton, 1987). In general, LFP has successfully detected between-group differences in production. However, Laufer (1998) found that LFPs were not predictive of learners' lexical progress, and thus are less successful in detecting pre-post, within-group differences. This could be a matter of using large units (word bands) to measure small distinctions (learner development).

The alternative to a grouping or band-based approach is a count-based approach (Tuldava, 1996), which involves averaging individual word frequencies to one-integer frequency ratings for texts. A simple version of this would be as follows: *The* (BNC frequency = 6,041,234) *cat* (3844) *sat* (11,038) *on* (729,518) *the* (6,041,234) *mat* (569) works out to an average word frequency value for the text of 2,137,906.17 (SD = 3,036,494.47). In contrast, the text The (6,041,234) lizards (196) basked (47) in (1,937,819) sunshine (629) on (729,518) igneous (129) rocks (2864) amounts to a word frequency value for the text of only 1,089,054.50 (SD = 2,114,424.55) or about half that of the first text. So by this measure the lexis of the second text is twice as rich as that of the first, as seems to correspond roughly to intuition. These numbers can be made more manageable either by calculating frequencies on a per-million basis (as is done in the official BNC lists by Leech et al., 2001) or by applying a mathematical function such as a logarithmic transformation to normalize the frequency distribution (i.e., a logarithmic transformation to the 10th for 6,041,234 equates to 6.78). The major advantage of a count-based approach is its closeness to the frequency data (i.e., there are no assumptions in count-based approaches that words should be grouped into families or that frequencies should be arbitrarily split into bands of 1000 word families). This closeness should afford greater accuracy in assessing lexical sophistication and measuring progress in learner production, thus providing some pedagogical rationale for employing such indices. For instance, a learner who can produce *gone* (18,433 BNC hits) should be assessed as having more lexically developed and sophisticated knowledge of English than one who can merely recognize it as a form of *go* (87,021 hits as an individual word form). As a result, count-based indices are almost certainly more likely to pick up small changes in learner development than band-based indices, because development over short periods would

inevitably occur mainly within rather than between bands (Horst and Collins, 2006). Evidence for the accuracy of count-based indices can be found in several recent studies in which count-based frequency indices have been predictive of human judgments of lexical proficiency (Crossley et al., 2011a, 2011b), human judgments of L2 writing quality (Crossley and McNamara, 2012), standardized proficiency test levels (i.e., TOEFL and ACT-Compass scores; Crossley et al., 2012), and lexical production at the early stages of L2 acquisition (Crossley et al., 2010).

However, count-based indices are not without their own limitations. A potential disadvantage of a count-based approach is that the measure is not purely lexical in that it captures morpho-syntactic as well as lexical knowledge. For instance, in our example above, the differences in word frequency between *go* and *gone* result mostly from morphological differences and not lexical differences. Additionally, count-based indices may be biased to productive properties of texts rather than receptive, the latter being arguably more important for early language learning. Lastly, count-based indices provide little intuitive information about the relevance of the values they report. While band-based indices can indicate, for example, the percentage of level-one words in a text, which allows interpretation by language practitioners, count-based indices provide numeric values that represent an entire text or textual elements (i.e., average minimum word frequency in sentences) and are difficult if not impossible to interpret.

In sum, the band-based approach has a track record with practitioners, but is somewhat biased to receptive knowledge and may be inaccurate in measuring productive knowledge. By contrast, the count-based approach, though potentially better at measuring productive knowledge, provides a less intuitive measure that practitioners may have difficulty interpreting. With these differences in mind, our research questions for this study are then as follows:

1. Can frequency based analyses of learner production predict language proficiency levels based on standard language tests?
2. Do band-based or count-based frequency analyses of learner productions predict these proficiency levels better?
3. When performing frequency analysis on texts, is language research concerning lexical proficiency better served by using band-based frequency analyses, count-based frequency analyses, or both in combination?

2. Methods

Our purpose of this study is to assess the effectiveness of automated and semi-automated indices of word frequency (both band-based and count-based) to distinguish between written samples produced by writers at a variety of proficiency levels (i.e., native speakers and L2 writers at beginning, intermediate, and advanced proficiency levels). Our goal is to assess the strengths of count-based and band-based frequency indices to classify learner productions based on the proficiency level of the writer. Such an analysis provides us with the opportunity to not only assess the reliability of frequency indices, but to better understand how word frequency in learner texts differs as a function of proficiency level.

2.1. Corpus construction

For our corpus, we collected unstructured writing samples (referred to henceforth as freewrites). These freewrites were unrelated to essay development (i.e., the freewrites were not used as a precursor to essay writing). Rather, the participants were asked to write about a topic of their choosing for 15 min. We selected freewrites so that both genre and topic expectations did not control the lexical output of the students. Freewrites such as these also involve more natural production and, thus, should better reflect the writers' lexical knowledge. We collected freewrites from 100 L2 learners using a cross-sectional approach. By level, there were 37 beginning level freewrites, 29 intermediate level freewrites, and 33 advanced level freewrites. We also randomly selected 30 native speaker freewrites from the Stream of Consciousness Data Set from the Pennebaker Archive Project (Newman et al., 2008). In the case of the L2 freewrites, the participants handwrote their freewrites and these were later converted to electronic text. The L1 freewrites were collected electronically. All of the L2 participants were studying English in the United States at one of two intensive language programs. The L2 learners came from 19 different L1 backgrounds (Arabic, Bambara, Bangla, Chinese, Dutch, French, German, Hindi, Ibo, Japanese, Korean, Farsi, Portuguese, Russian, Spanish, Thai, Turkish, Vietnamese, and Yoruba) and ranged in age from 17 to 34 years old. The L1 freewrites were collected from college freshmen at a university in the United States.

We controlled for text length in the freewrite samples by selecting text segments of about 150 words from each sample. This process was done randomly and was based on paragraph constraints (i.e., text samples were separated at the paragraph level and not at the word or sentence level). We used an automatic spellchecker to correct all samples and an L1 speaker then confirmed the corrections. Table 1 shows average text sizes and number of participants at each proficiency level.

2.2. Level classification

We classified all L1 writers as native speakers of English. We used two different proficiency tests to classify the L2 writers as beginning, intermediate, or advanced speakers of English. As part of the evaluation and placement for their respective intensive English programs, each participant was administered either the TOEFL (either the internet-based test or the Institutional TOEFL) or the ACT Compass computer-adaptive ESL reading and grammar tests.

We used the total score on both the internet-based test or the institutional versions of the TOEFL exams to classify the L2 writing samples. We used proficiency categories (beginner, intermediate, and advanced) suggested by Wendt and Woo (2009) and Boldt et al. (1992) to classify the TOEFL participants. We could find no direct comparisons between the TOEFL tests and the ACT ESL Compass test. Thus, we used the test maker's suggested descriptors and proficiency levels to classify the freewrites of the ACT participants. We used the following classifications for the L2 writers: a score of 126 or below on the combined Compass ESL reading/grammar tests, 32 or below on the TOEFL iBT, or 400 or below on the TOEFL PBT classified the participant as a beginning level L2 learner. A score between 127 and 162 on the combined Compass ESL reading/grammar tests, 33 and 60 on the TOEFL iBT, or 401 and 499 on the TOEFL PBT classified the participant as an intermediate level L2 learner. A score of 163 and above on the combined Compass ESL reading/grammar tests, 61 and above on the TOEFL iBT, or 500 or above on the TOEFL PBT classified the participant as an advanced level L2 learner. Such classifications have been used in similar studies concerning lexical proficiency (Crossley et al., 2011a, 2011b; 2012).

2.3. Selected frequency indices

To examine the frequency of the words in each freewrite, we used both band-based and count based frequency indices. The band-based indices were LFP (Laufer and Nation, 1995; but BNC-adapted in line with Nation, 2006), and P_Lex (Meara and Bell, 2001); the count-based indices were computed from the CELEX frequency norms (Baayen et al., 1995). The LFP indices were collected from the website http://www.lextutor.ca/vp/bnc/. P_Lex indices were collected from the semi-automated tool P_Lex v2.0 (available at http://www.lognostics.co.uk/). The CELEX frequency indices were reported by Coh-Metrix (available at http://cohmetrix.com/). The selected indices are discussed below.

2.3.1. Lexical Frequency Profiles (LFP)

The version of LFP used for this study reports word frequency for twenty 1000-word family bands, moderated by range and genre information, as described in the Introduction. Forty-two separate frequency indices for each freewrite were computed from LFP. Twenty indices corresponded to the percentage of word at each k-level (i.e., the percentage of level one words in the text, the percentage of level two words in the text). Twenty indices reflected the total percentage of words accounted for at each level (i.e., the percentage of words in the text accounted for at level 10 or the percentage of words reported between level 1 and level 10). The final index in this categorization (the percentage of words in the text accounted for between levels 1 and 20), inversely corresponds to the number of off list words (e.g., if 95% of the words in a text are accounted for at level 20, then 5% of the words in the text are off list words). Lastly, two additional indices measure the level at which the freewrite contained 95% and 98% of all the words in the text, the

Table 1
Descriptive statistics for freewrite corpus.

Level	Participants	Mean number of words	Number of words standard deviation
Beginning	37	156.946	40.055
Intermediate	29	177.433	51.541
Advanced	33	162.455	27.621
Native speaker	30	139.900	16.076

percentage of words that studies have shown corresponds to text comprehensibility (95% according to Laufer, 1992, and 98% according to Nation, 2006).

2.3.2. P_Lex

P_Lex calculates the frequency of the words in a text by computing the Poisson distributions (i.e., sensitivity to infrequent events in a long series of trials) of difficult words in ten word text segments. To do this, P_Lex divides a text into 10 word segments ignoring punctuation or sentence boundaries. Using Nation's original word lists (those used in Laufer and Nation, 1995), P_Lex then calculates the number of easy words per 10 word segments, with easy words defined as words that occur in the first band of Nation's word list (i.e., the 1000 most frequent words and their derivatives) along with all proper nouns, numbers, and geographical derivatives. All other words are categorized as difficult or infrequent. If P_Lex does not recognize a word, it prompts the user to classify the word as easy or difficult (as a result P_Lex is not truly automatic). The P_Lex profile for a text is the number of segments containing zero infrequent words, the number containing one infrequent word, the number containing two infrequent words, and so forth. The distribution of these words is Poisson. These Poisson distributions are then converted to a Lambda value. Lambda values for P_Lex range from 0 to 45, with higher scores reflecting a text containing more infrequent words. Meara and Bell (2001) argue that Lambda values are less sensitive to text length than other band-based measures such as LFP. Two lambda indices are reported by P_Lex: lambda values including easy or first-1000 level words (level 0 words) and lambda values excluding easy words.

2.3.3. CELEX norms

The CELEX word frequency measurements comprise frequencies computed from the early 1991 version of the COBUILD corpus, a 17.9 million word corpus (over 50,000 word types; Baayen et al., 1995). The text analysis package Coh-Metrix reports the means, standard deviations, and minimum frequency values for content words only as well as for all words in a sample text, based on the CELEX norms. The mean scores are reported both as raw values and as logarithmic values (to the logarithm of 10). Coh-Metrix also reports mean and standard deviation values for the average word frequency at the level of the sentence, the paragraph, and the text. The mean values are based on the average frequency values for each word. The standard deviation scores are based on the average standard deviations between the words. These values can be averaged across sentence and paragraphs or reported as an average for the entire text. Lastly, Coh-Metrix reports frequency values taken from the entire COBUILD corpus (including both written and spoken subset corpora), for the spoken subset corpus contained in COBUILD (which consists of 1.3 million spoken tokens), and for the written subset corpus (the remaining 16.6 million words). In total, this combination of parameters leads to just over 70 different CELEX indices reported by Coh-Metrix, all of which we test in this study. So, for instance, Coh-Metrix reports a value for the logarithmic mean of the CELEX frequency values for only the content words in a sample text averaged across sentences that are found in the spoken part of the COBUILD Corpus. For the average and minimum frequency values, a lower value equals less frequent words while a higher value equals more frequent words. For instance, the frequent word *think* has a value of 3.3 in the total corpus, 3.19 in the written corpus, and 3.87 in the spoken corpus while the less frequent word *consider* has a value of 2.31 in the entire corpus, 2.32 in the written corpus, and 2.13 in the spoken corpus. Thus a sample text that reports a logarithmic mean frequency value of 2.504 contains less frequent words than a sample text that reports a logarithmic mean frequency value of 2.699. For indices based on standard deviation, a lower standard deviation in frequency scores indicates that the words in a text tend to have frequency values closer to one another than the frequency of words in a text that have higher standard deviations.

2.4. Statistical analysis

For each set of frequency indices (i.e., LFP, P-Lex, CELEX), we first conduct a Multiple Analysis of Variance (MANOVA) to test if the reported indices in each set demonstrate significant differences between the freewrites according to proficiency levels. Next, we conduct a stepwise discriminant function analysis (DFA) using only the indices from each set that showed significant differences between the proficiency levels, but did not exhibit multicollinearity with other indices in the set.[2] A discriminant function is generated by the DFA. This discriminant function

[2] Multicollinearity between indices indicates that the indices are effectively measuring the same patterns in the data.

Table 2
Means (standard deviations) for LFP values and text levels.

Variables	Beginner	Intermediate	Advanced	Native speaker	f (3, 126)	p	η_p^2
Percentage of words between levels one and twenty	95.628 (3.710)	96.158 (2.976)	97.050 (2.497)	98.374 (1.452)	5.878	<0.001	0.123
Level at which ninety five percent of words in text are used	1.865 (0.631)	2.133 (0.681)	2.576 (1.0316)	2.433 (1.040)	4.681	<0.010	0.100
Percentage of level four words	0.445 (0.666)	0.447 (0.830)	0.895 (0.920)	1.035 (0.925)	4.291	<0.050	0.093
Percentage of level five words	0.352 (0.501)	0.424 (0.703)	0.892 (1.055)	0.534 (0.744)	3.245	<0.050	0.072
Percentage of level two words	3.938 (2.439)	5.069 (2.958)	5.750 (2.678)	4.550 (2.586)	2.887	<0.050	0.064
Percentage of level ten words	0.069 (0.206)	0.023 (0.124)	0.072 (0.258)	0.191 (0.324)	2.747	<0.050	0.061

η_p^2 = Partial eta squared.

produces an algorithm that can be used to predict group membership (i.e., the proficiency level of the writers). Finally, we conduct a combined analysis using all the selected indices from LFP, P-Lex, and CELEX to see which were most predictive of proficiency level.

We first conduct a DFA on the entire set of freewrites. The model reported by this DFA is then used to predict group membership of the freewrites using leave-one-out-cross-validation (LOOCV). In this type of validation, a fixed number of folds equal to the number of observations (i.e., the 130 texts) is selected. For each fold, one observation in turn is left out and the remaining instances are used as the training set (in this case the 129 remaining freewrites). We test the accuracy of the model based on its ability to predict the proficiency classification of the omitted instance. The LOOCV procedure allows testing of the accuracy of the model on an independent data set (i.e., on data that is not used to train the model). If the discriminant analysis model for both the entire set and the *n*-fold cross-validation set predict similar classifications, then the strength of the model to extend to external data sets is supported.

Our comparison of the different frequency indices thus focuses on the classifications of the texts as categorized by the proficiency level of the writers and the predictions made by the DFA model. To illustrate these comparisons, we report the results in terms of precision, recall, and F1 score. We computed recall scores by tallying the number of hits (correct predictions) over the number of hits + false negatives (i.e., the number of beginning level freewrites that were misclassified as intermediate, advanced or native speaker). Thus if there are 40 freewrites at the beginning level and the algorithm gives 30 correct predictions and 10 incorrect predictions, the recall score is $(30/(30 + 10)) = 75\%$. Precision is the number of hits divided by the number of hits + false positives (i.e., the number of intermediate, advanced, and native speaker freewrites that were classified as beginning level). Thus, if 30 beginning freewrites are correctly predicted as beginning level and 10 intermediate and 10 advanced freewrites are also categorized as beginners, the precision score is $(30/(30 + 10+10)) = 60\%$. Reporting both precision and recall allows us to understand to a greater degree the accuracy of the model reported by the DFA. The F1 score is basically a weighted average of the precision and recall results.

To summarize, we analyze each set of indices independently (i.e., the band-based LFP and P-Lex indices and the count-based CELEX indices) and combined. We first use a MANOVA to examine if the reported indices for the set demonstrate significant differences between the freewrites according to proficiency levels. We then use a DFA to test the hypothesis that the frequency indices can be used to accurately classify the freewrite proficiency levels. We then go on to compare differences in the accuracy of the index types (LFP, P_Lex, and CELEX) in classifying the freewrites in order to assess if one index type is superior to the other.

3. Results

3.1. Lexical Frequency Profiles

3.1.1. MANOVA

A MANOVA was conducted using the LFP indices as the dependent variables and the previously categorized proficiency ratings for the freewrites as the independent variables. All variables that reported significant differences were then assessed using Pearson correlations for multicollinearity (with a threshold of $r > 0.70$). Six out of 42 LFP-based indices demonstrated significant difference between proficiency levels while not demonstrating multicollinearity with one another. These variables were percentage of words between level one and twenty, level at which ninety-five percent of words in text were used, percentage of level four words, percentage of level five words,

Table 3
Confusion matrix of LFP indices: Predicted level versus actual level (total and cross-validated set).

Actual text type	Predicted text type			
	Beginner	Intermediate	Advanced	Native speaker
Total set				
Beginner	17	10	5	5
Intermediate	6	15	5	4
Advanced	8	7	13	5
Native speaker	2	3	8	17
Cross-validated set				
Beginner	15	11	5	6
Intermediate	7	12	7	4
Advanced	8	7	13	5
Native speaker	2	3	8	17

percentage of level two words, and *percentage of level ten words*. Descriptive statistics and MANOVA results for the six significant indices are presented in Table 2, ordered by effect size.

3.1.2. Discriminant function analysis

The stepwise DFA selected variables from Column 1 in Table 2 based on a statistical criterion that retains the variables that best classify the grouping variable (proficiency level) and helps control for potential multicollinearity. For our analysis, the significance level for a variable to enter or to be removed from the model was set at the norm generally adopted in applied linguistics: $p \leq 0.05$. The stepwise DFA retained four variables as significant predictors of proficiency level (percentage of words between levels one and twenty, level at which ninety-five percent of words in text were used, percentage of level ten words, and *percentage of level four words*) and removed the remaining two variables as non-significant predictors.

The results demonstrate that the DFA using the four significant LFP indices correctly allocated 62 of the 130 freewrites in the total set, χ^2 (df = 3, $n = 130$) = 43.583, $p < 0.001$, for an accuracy of 47.7% (the chance level for this analysis and all analyses is 25% because there are four proficiency groupings and hence a 1 in 4 chance of a chance positive classification).[3] For the leave-one-out cross-validation (LOOCV), the discriminant analysis correctly allocated 57 of the 130 freewrites for an accuracy of 43.8% (see the confusion matrix reported in Table 3 for results).[4] The measure of agreement between the actual text type and that assigned by the model produced a weighted Cohen's Kappa of 0.373, demonstrating a fair agreement.[5]

The precision scores (the ratio of hits to hits + false negatives) and recall scores (hits over hits + false positives) from the model for predicting the level of the freewrites using the LFP indices are presented in Table 4. The model performed best for native speaker freewrites and performed worst for intermediate L2 freewrites. The overall accuracy of the model for the total set was 0.478 (the average F1 score), and for the cross-validated set was 0.440. The results demonstrate that the combination of four LFP indices can discriminate between proficiency levels to a fair degree.

3.2. P_Lex

3.2.1. MANOVA

A MANOVA was conducted using the P_Lex indices as the dependent variables and the proficiency levels of the freewrites as the independent variables. All variables that reported significant differences were then assessed using

[3] The baseline percentages for this analysis are 29% for beginning texts, 23% for intermediate texts, 25% for advanced texts, and 23% for native speaker texts.

[4] A confusion matrix displays the number of correct and incorrect predictions made by a model. The first row in the confusion matrix in Table 3 demonstrates that 17 of the 37 beginning freewrites were correctly classified in the total set DFA. Ten beginning freewrites were misclassified as intermediate, five as advanced, and five as native speaker freewrites. The first column in Table 3 also shows that 17 of the 37 beginning freewrites were correctly classified in the total set DFA. In addition, the column shows that six intermediate, eight advanced, and two native speaker freewrites were misclassified as beginning level freewrites.

[5] Cohen's kappa coefficient is a statistical measure between 0 and 1 of inter-rater agreement for categorical items that incorporates an estimate of chance agreement.

Table 4
Precision and recall results for LFP indices (Total and cross-validated set).

Text set	Precision	Recall	F1
Total set			
Beginner	0.459	0.515	0.486
Intermediate	0.500	0.429	0.462
Advanced	0.394	0.419	0.406
Native speakers	0.567	0.548	0.557
Cross-validated set			
Beginner	0.405	0.469	0.435
Intermediate	0.400	0.364	0.381
Advanced	0.394	0.394	0.394
Native speakers	0.567	0.531	0.548

Pearson correlations for multicollinearity (with a threshold of $r > 0.70$). Only one index out of two demonstrated significant difference between proficiency levels: *lambda value without level 0 words* (i.e., without the first 1000-level "easy" words). Descriptive statistics and MANOVA results for this index ordered by effect size are presented in Table 5.

3.2.2. Discriminant function analysis

The stepwise DA retained the variable *lambda value without level 0 words* as a significant predictor. The results demonstrate that the DFA using the one index correctly allocated 47 of the 130 freewrites in the total set, χ^2 (df = 3, $n = 130$) = 26.867, $p < 0.001$, for an accuracy of 36.2%. For the cross-validated set, the discriminant analysis correctly allocated 41 of the 130 freewrites for an accuracy of 31.5% (see Table 6 for results). The measure of agreement between the actual text type and that assigned by the model produced a weighted Cohen's Kappa of 0.305, demonstrating a fair agreement.

Table 7 provides the P_Lex index precision and recall scores for predicting the level of the freewrites. The model performed best for beginning L2 freewrites and worst for advanced L2 freewrites. The overall accuracy of the model for the total set was 0.341 (the average F1 score). The accuracy for the cross-validated set was 0.296. The results demonstrate that the one P_Lex index can discriminate between proficiency levels to a fair degree.

3.3. CELEX norms

3.3.1. MANOVA

A MANOVA was conducted using the CELEX indices as the dependent variables and the freewrites as the independent variables. All variables that reported significant differences were then assessed using Pearson correlations for multicollinearity (with a threshold of $r > 0.70$). Nine indices out of the 70 CELEX indices demonstrated significant difference between proficiency levels while not demonstrating multicollinearity with one another. These variables were Minimum of CELEX content word spoken frequency by logarithm in sentence, Minimum of CELEX content word written frequency by logarithm in paragraph, Minimum CELEX content word written frequency in paragraph, Standard deviation of CELEX content word spoken frequency in sentence, CELEX content word frequency by logarithm in sentence, CELEX spoken frequency by logarithm in paragraph, Minimum CELEX content word written frequency in sentence, Standard deviation of CELEX content word spoken frequency by logarithm in sentence, and *Standard deviation of CELEX spoken frequency in sentence*. Descriptive statistics and MANOVA results for these indices ordered by effect size are provided in Table 8.

3.3.2. Discriminant function analysis

The stepwise DA retained four variables as significant predictors (Minimum of CELEX content word spoken frequency by logarithm in sentence, Minimum CELEX content word written frequency in paragraph, CELEX spoken

Table 5
Means (standard deviations) for P_Lex value and text levels.

Variables	Beginner	Intermediate	Advanced	Native speaker	f (3, 126)	p	η_p^2
Lambda value without level 0	0.871 (0.488)	1.146 (0.514)	1.602 (0.726)	1.618 (0.661)	12.426	<0.001	0.228

Table 6
Confusion matrix for P-Lex Lambda value without level 0: Predicted level versus actual level (total and cross-validated set).

Actual text type	Predicted text type			
	Beginner	Intermediate	Advanced	Native speaker
Total set				
Beginner	21	10	3	3
Intermediate	13	7	4	6
Advanced	5	11	5	12
Native speaker	6	3	7	14
Cross-validated set				
Beginner	21	10	3	3
Intermediate	13	7	4	6
Advanced	5	11	5	12
Native speaker	6	3	13	8

frequency by logarithm in paragraph, and *Standard deviation of CELEX spoken frequency in sentence*) and removed the remaining five variables as insignificant predictors.

The results demonstrate that the DFA using the four CELEX indices correctly allocated 75 of the 130 freewrites in the total set, χ^2 (df = 3, n = 130) = 77.617, $p < 0.001$, for an accuracy of 57.7%. For the cross-validated set, the discriminant analysis correctly allocated 62 of the 130 freewrites for an accuracy of 47.7% (see Table 9 for results). The measure of agreement between the actual text type and that assigned by the model produced a weighted Cohen's Kappa of 0.452, demonstrating a moderate agreement.

The precision and recall scores from the model for predicting the level of the freewrites using the CELEX indices are presented in Table 10. The model performed best for beginning L2 freewrites and worst for advanced L2 freewrites. The overall accuracy of the model for the total set was 0.574 (the average F1 score). The accuracy for the cross-validated set was 0.471. The results demonstrate that the combination of four CELEX indices discriminate between proficiency levels to a moderate degree.

3.4. Combined analysis

3.4.1. Variable selection

To assess the classification strength of band-based and count-based frequency indices as a whole, we conducted a DFA that combined the significant band-based indices (LFP and P_Lex; see Tables 2 and 5) and the count-based indices (the CELEX indices; see Table 8). We first assessed multicollinearity between the 16 variables using Pearson correlations. None of the variables demonstrated correlations of $r > 0.70$.

3.4.2. Discriminant function analysis

The stepwise DFA retained three CELEX variables (Minimum of CELEX content word spoken frequency by logarithm in sentence, CELEX spoken frequency by logarithm in paragraph, and *Minimum of CELEX content word written frequency by logarithm in paragraph*), but no LFP or P_Lex variables.

Table 7
Precision and recall results for P-Lex Lambda value without level 0 (Total and cross-validated set).

Text set	Precision	Recall	F1
Total set			
Beginner	0.568	0.467	0.512
Intermediate	0.233	0.226	0.230
Advanced	0.152	0.263	0.192
Native speakers	0.467	0.400	0.431
Cross-validated set			
Beginner	0.568	0.467	0.512
Intermediate	0.233	0.226	0.230
Advanced	0.152	0.200	0.172
Native speakers	0.267	0.276	0.271

Table 8
Means (standard deviations) for CELEX values and text levels.

Variables	Beginner	Intermediate	Advanced	Native speaker	$F(3, 126)$	p	η_p^2
Minimum of CELEX content word spoken frequency by logarithm in sentence	1.788 (0.281)	1.573 (0.312)	1.428 (0.264)	1.335 (0.354)	14.561	<0.001	0.257
Minimum of CELEX content word written frequency by logarithm in paragraph	0.859 (0.652)	0.464 (0.453)	0.410 (0.424)	0.103 (0.272)	14.031	<0.001	0.25
Minimum CELEX content word written frequency in paragraph	494.116 (850.498)	103.833 (158.142)	71.303 (104.597)	16.800 (36.686)	7.746	<0.001	0.156
Standard deviation of CELEX content word spoken frequency in sentence	5877.792 (2930.873)	4506.311 (1171.060)	4732.328 (1348.947)	3814.405 (1751.803)	6.336	<0.001	0.131
CELEX content word frequency by logarithm in sentence	2.678 (0.182)	2.600 (0.182)	2.573 (0.153)	2.509 (0.174)	5.476	<0.001	0.115
CELEX spoken frequency by logarithm in paragraph	2.761 (0.218)	2.924 (0.166)	2.836 (0.157)	2.869 (0.139)	5.103	<0.010	0.108
Minimum CELEX content word written frequency in sentence	6677.959 (1286.805)	2011.168 (1785.018)	1263.174 (806.303)	1607.543 (2382.504)	4.629	<0.010	0.099
Standard deviation of CELEX content word spoken frequency by logarithm in sentence	0.501 (0.114)	0.433 (0.120)	0.413 (0.102)	0.411 (0.160)	4.064	<0.010	0.088
Standard deviation of CELEX spoken frequency in sentence	4026.501 (1001.917)	4197.059 (1238.106)	3383.446 (1088.839)	3582.224 (1080.559)	3.797	<0.050	0.083

The results using the three CELEX variables show that the DFA correctly allocated 69 of the 130 freewrites in the total set, χ^2 (df = 3, n = 130) = 67.346, $p < 0.001$, for an accuracy of 53.1%. For the cross-validated set, the discriminant analysis correctly allocated 64 of the 130 freewrites for an accuracy of 49.2% (see Table 11 for results). The measure of agreement between the actual text type and that assigned by the model produced a weighted Cohen's Kappa of 0.420, demonstrating a moderate agreement.

The precision and recall scores from the model for predicting the level of the freewrites using the three CELEX indices are presented in Table 12. The model performed best for beginning L2 freewrites and worst for advanced L2 freewrites. The overall accuracy of the model for the total set was 0.526 (the average F1 score). The accuracy for the cross-validated set was 0.487. The results demonstrate that the combination of the three CELEX indices discriminate between proficiency levels to a moderate degree.

Table 9
Confusion matrix for CELEX indices: Predicted level versus actual level (total and cross-validated set).

Actual text type	Predicted text type			
	Beginner	Intermediate	Advanced	Native speaker
Total set				
Beginner	24	4	4	5
Intermediate	2	16	7	5
Advanced	6	3	17	7
Native speaker	3	5	4	18
Cross-validated set				
Beginner	23	4	5	5
Intermediate	4	13	7	6
Advanced	6	4	13	10
Native speaker	3	7	7	13

Table 10
Precision and recall results for CELEX indices (Total and cross-validated set).

Text set	Recall	Precision	F1
Total set			
Beginner	0.649	0.686	0.667
Intermediate	0.533	0.571	0.552
Advanced	0.515	0.531	0.523
Native speaker	0.600	0.514	0.554
Cross-validated set			
Beginner	0.622	0.639	0.630
Intermediate	0.433	0.464	0.448
Advanced	0.394	0.406	0.400
Native speaker	0.433	0.382	0.406

3.5. Comparisons between frequency indices

We assigned each freewrite either a 0 or a 1 based on whether the frequency index had accurately predicted its group membership (0 = inaccurate, 1 = accurate) in the total set analysis. Table 13 provides the means and standard deviations where perfect predictive ability would be reflected by a mean score of 1. We conducted *t*-tests between the classification results for each frequency index to assess the significance of differences in classification accuracy existed between the indices. To control for Type 1 errors (i.e., false positives), we used a Bonferroni Correction and lowered our criterion for significance to $p = 0.015$. No significant differences in classification accuracy were reported between LFP indices and P_Lex, $t(258) = -1.891$, $p = 0.060$, or the LFP and CELEX indices, $t(258) = -1.617$, $p = 0.107$. Significant differences were reported between the CELEX indices and P_Lex, $t(258) = -3.550$, $p < 0.001$. The results demonstrate that the predictions made by the CELEX indices were significantly more accurate than those made by P_Lex, but that the differences in accuracy between LFP indices and P_Lex indices and the differences between CELEX indices and LFP indices were not significant.

4. Discussion and conclusion

This study has demonstrated that the frequency indices reported by LFP, P_Lex, and Coh-Metrix can significantly predict the proficiency level classification of texts to varying degrees. The highest success rate was reported for the CELEX indices computed by Coh-Metrix (58% accuracy). These indices reported a moderate Kappa value between the human classification and the classification derived from the discriminant analysis. This classification accuracy was significantly better than that reported by P_Lex, but not significantly better than that reported by the LFP indices. The LFP indices reported the second highest classification accuracy (48% accuracy) with a fair agreement between the actual classification and the classifications reported by the discriminant analysis (according to the Kappa value). However, the LFP indices did not report significantly higher classification rates than either the CELEX or P_Lex indices. Our weakest predictor was the P_Lex indices, which accurately classified only 36% of the texts based on

Table 11
Confusion matrix for combined indices: Predicted level versus actual level (total and cross-validated set).

Actual text type	Predicted text type			
	Beginner	Intermediate	Advanced	Native speaker
Total set				
Beginner	24	7	3	3
Intermediate	0	16	7	7
Advanced	6	8	12	7
Native speaker	2	8	3	17
Cross-validated set				
Beginner	24	7	3	3
Intermediate	0	15	7	8
Advanced	6	8	11	8
Native speaker	2	8	6	14

Table 12
Precision and recall results for combined indices (Total and cross-validated set).

Text set	Recall	Precision	F1
Total set			
Beginner	0.649	0.75	0.696
Intermediate	0.533	0.41	0.464
Advanced	0.364	0.48	0.414
Native speaker	0.567	0.5	0.531
Cross-validated set			
Beginner	0.649	0.750	0.696
Intermediate	0.500	0.395	0.441
Advanced	0.333	0.407	0.367
Native speaker	0.467	0.424	0.444

proficiency level. Such accuracy demonstrated only fair agreement between actual and predicted classification. The P_Lex classifications were significantly less accurate than the CELEX classifications, but no less accurate than the LFP indices. A combined analysis using LFP, P_Lex, and CELEX indices retained only CELEX indices and removed the LFP and P_Lex indices. This analysis reported a classification accuracy of 52%.

Of interest in the above models are instances where native speakers (NSs) were classified as non-native speakers (NNSs) and vice-versa. In eight cases there was a clear pattern of misclassification wherein 75% of the models (the LFP, P_Lex, CELEX, and combined models) misclassified the participant as being a NS or NNS. Half of these cases ($n = 4$) involved NNSs classified as NSs. In no case, however, was a beginning level NNS misclassified as a NS. In two cases, intermediate NNSs were classified as NSs and in two cases advanced NNSs were classified as NSs. The other half of the cases ($n = 4$) involved NSs misclassified as NNSs. In three of these cases, 75% of the models classified the NSs as beginning or intermediate level NNSs. Overall, these misclassifications demonstrate that intermediate and advanced NNSs can produce enough infrequent words that they can misclassified as NSs and that some NSs produce enough frequent words to be misclassified as NNSs.

In general, all the frequency indices indicated that as L2 learners advance in proficiency level, they begin to use less frequent words, with native speakers of English using the least frequent words. The LFP indices, for example, indicated that beginning level L2 learners produced 95% of the words in the text at a developmentally earlier level than advanced L2 learners and native speakers. Thus, the majority of the words produced by beginning level writers were produced at a lower level than advanced L2 and native speaker writers. This result provides further evidence that L2 learners, like L1 learners (Biemiller and Slonim, 2001), tend to demonstrate frequency patterns that develop concomitantly with proficiency. Beginning level L2 learners also produced a lower percentage of level two, level four, level five, and level ten words than advanced L2 learners and native speakers. These indices indicate the degree to which the text deploys words that go beyond the language of speech and conversation (Adolphs and Schmitt, 2003), and reinforce the importance of "mid-frequency vocabulary" in L2 development (Schmitt and Schmitt, 2013). The P-Lex index also demonstrated that beginning level L2 learners produced more frequent words across the texts, while the Coh-Metrix indices demonstrated that beginning level L2 learners produced more frequent words at the sentence and paragraph level.

The frequency indices tested in this study also demonstrated other properties of lexical development beyond straight frequency counts. For instance, the LFP index *percentage of words between levels one and twenty* showed that as L2 learners advanced they produced more words within the first 20 bands of English with beginning learners' freewrites containing 95.628% of tokens within these bands and advanced learners' freewrites containing 97.050% (and native speakers' freewrites containing 98.374%). At first glance, this finding may seem counterintuitive, but the results indicate that beginning level writers produce more off-list words (proper nouns, non-standard morphologies, and non-words) than advanced L2 learners and native speakers, who tend to use a higher proportion of standard lexical words.

Table 13
Descriptive statistics for *t*-test data.

Frequency index	Mean	Standard deviation	*N*
LFP	0.477	0.501	130
P_Lex	0.362	0.482	130
CELEX	0.577	0.496	130

The standard deviation indices reported by Coh-Metrix based on the CELEX norms are also of interest. These indices report similarities in word frequency use across a text such that lower values indicate that writers produce words of a similar frequency while higher values indicate that writers produce words with a greater range of frequencies. These indices all demonstrated that beginning level writers produced words that had a greater range of frequencies than advanced L2 learners and native speakers, who tended to use words that were closer in frequency to one another. Such a finding indicates that the beginning level L2 learners lacked consistency in the frequency of the words they produced. Examples of this can be seen in three excerpts from beginning level L2 writers:

1. I have many friends in my class. My classmates are very kindly and funny.
2. My family is big family. I have one elder sister and two younger sister.
3. I take care two children. My pet is clever dog.

These examples demonstrate that beginning level L2 learners will often use infrequent words (*kindly*, *elder*, *clever*) where more frequent words would be expected. The use of infrequent words interspersed within a sample containing mostly frequent words will lead to the high standard deviations reported by the CELEX indices. The use of such infrequent words likely stems from underdeveloped lexicons that contain few synonyms or grammars without set morphological rules (in the case of the word *kindly*).

A further point of possible interest is the power of spoken CELEX frequencies as predictors of proficiency. This power could reflect the naturalistic quality of the freewrites examined and the potential for these freewrites to contain more conversational language if compared to texts written on assigned topics. In total, five of the nine CELEX indices that demonstrated differences between the proficiency levels were spoken indices (see Table 8) including the index that demonstrated the strongest relationship with the proficiency level of the freewrite (*Minimum of CELEX content word spoken frequency by logarithm in sentence*). Such findings may indicate that the freewrites used in this analysis are more representative of naturalistic lexical production because their levels are best predicted using indices derived from spoken discourse. This may not be the case for corpora based on essays as used in previous studies (Laufer and Nation, 1995).

In reference to our research questions, the answer to the first research question (Can frequency based analyses of learner production predict scores on standard language tests?) appears to be affirmative, at least in principle. The results indicate that indices from each of the frequency measures we tested in this study were able to predict to a significant degree the proficiency level of the writers for the freewrites. The levels of accuracy ranged from 32% to 58% indicating fair to moderate agreement with the actual classification of the writers' proficiency levels. Unlike past studies (i.e., Laufer and Nation, 1995), ours had stronger controls for proficiency level because we relied on standardized language assessment tests. We also collected writing samples from L2 learners from 19 different L1 backgrounds. These criteria give us greater confidence that the differences in our groups are based on proficiency level and that the frequency indices assessed in this study are generalizable to a wide variety of learners. Additionally, the statistical methodology we employed allowed us to run confirmatory machine learning models that classified the writing samples based on proficiency level as compared to solely examining statistical differences in means between levels as reported by an ANOVA analysis (cf. Laufer and Nation, 1995) or independent *t*-tests (Meara and Bell, 2001). Thus, we have confidence that the differences reported between the groups are not only significant but can be meaningfully applied to accurately classify texts.

The answer to the second research question (whether band-based or count-based frequency analyses better predict writing scores) appears to be that the count-based measures are better for this purpose. The count-based indices reported by Coh-Metrix reported moderate agreements between the actual classification and the predicted classification while the agreements for the band-based indices were only fair. The count-based indices also reported higher classification accuracy rates, although these rates were significantly higher only when compared to the P_Lex values, not the LFP values. In addition, when a combined analysis was conducted that included band-based and count-based indices, only the count-based indices were retained in the DFA indicating that the band-based indices did not explain a significant amount of variance beyond that explained by the count-based indices.

The third research question concerns whether we need two kinds of frequency analysis or one. For learner productions, the count-based approaches examined above seem a necessary addition to future applied linguistics toolkits, and it thus behooves the originators of these methods to begin work on the generation of norms for a variety of different classifications (e.g., What are the typical count-based values for spoken and written texts? What are the typical count-based values for texts written by beginner, intermediate, advanced L2 learners? What are the typical count-based values for simplified and authentic reading texts?). And yet there are important limitations to the count-based indices. One is

their lack of interpretability and the problem of deriving pedagogical information from them. It is difficult to say what advice to give learners who do not perform well on a count-based frequency measure. Presumably the implicit advice to the learner is to try to learn and to use less frequent words, and less frequent morphologies and derivations of words they already know. But which words? This remains unspecified and ungeneralizable in a count-based measure. Learners cannot be told to simply go and learn infrequent items, especially if, as in the CELEX indices, these have not been controlled for range or genre as LFP's lists have. A mid- or low-frequency item that appears only in one small part of a corpus will not normally be worth learning for a non-specialist learner. On the other hand, the LFP values come attached to a program and pedagogically relevant frequency lists that can be incorporated directly into curricula.

Here we provide an extended three-step example of a band-based individual intervention with both empirical support and an ongoing application. We consider a typical scenario where an academic ESL learner is having trouble reading his course texts. As a result, his instructor asks him to take Beglar and Nation's (2007) BNC thousand-family bands Vocabulary Size Test, with results showing competence up to and including the second thousand level but weakness thereafter. The instructor runs samples of the student's reading materials through the Lexical Frequency Profiler and determines that the texts this student is attempting to read contain an average 10% third-through-fifth thousand level items (as indeed is typical in academic texts). Or, from another angle, the student's 2000 available word families are providing him with only about 75% known-word coverage in his readings, while at least 95% known-word coverage has been shown to be needed (Laufer, 1992) – which knowing 5000 families would assure. As a result, the instructor assigns the student the task of going through each of the third, fourth and fifth-thousand level frequency lists, roughly one per week, and building an informal dictionary of those items the student does not already know. This can be done on Lextutor at www.lextutor.ca/list_learn/, where every item in 20 BNC lists is connected to a range of corpora from which to mine examples, a multilingual dictionary, a text-to-speech routine, links to several game-like practice activities, and a means of quickly assembling many of these components into an Excel-friendly glossary. This technology for scoping a set of frequency bands was piloted in Cobb (1997), validated in Cobb (1999), and, since being configured as an Internet program in 2006, has been used by thousands of instructors and independent learners. The whole scheme, however, depends on the availability of a coverage-validated, band-based frequency scheme, and it is not easy to imagine how a comparably targeted and rapid vocabulary expansion scheme could be developed within a count-based account of word frequency.

Thus, both methods have their advantages and, just as it is common enough for different grain sizes to be needed for different parts of a task (like cooking a dinner, where cup measures are adequate for making stew but half-teaspoons needed for the baking soda in cake), we suggest the incorporation of both methods. For instance, one can easily imagine a language learning situation where learners are placement tested with a band-based measure (Meara and Buxton, 1987; Beglar and Nation, 2007), enter a course of study comprising authentic materials that are either selected by a count-based frequency measure (Crossley et al. 2011, 2008) or adapted by a band measure (Cobb, 2007), or both, all supported by a band-based vocabulary course matched to placement level, with progress tallied at the end by a count-based measure that picks up small differences in lexical deployment.

5. Conclusion

In conclusion, this study has demonstrated the strength of count-based and band-based indices to distinguish texts categorized by the proficiency level of the writers. We find that count-based indices are stronger predictors of proficiency level, but note that the output produced by these indices may not lead to salient pedagogical applications. Additionally, it is not clear if count-based indices measure strictly vocabulary or something more akin to lexico-grammaticality wherein the indices measure a combination of word frequency and morphology. Lastly, future studies should investigate the pedagogical applications of count-based indices and assess the linguistic features contained within the values they report. Only when we have a better understanding of the strengths and weaknesses of count based indices can we make final assessments of how these indices can supplement or complement band-based indices.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors would also like to thank the anonymous reviewers and the editors and staff of System for their

support. Lastly, the authors would like to thank Scott Jarvis and Michael Daller for inviting them to the colloquium The validity of vocabulary measures at the 2011 American Association for Applied Linguistics conference from which the ideas in this paper derive.

References

Adolphs, S., Schmitt, N., 2003. Lexical coverage of spoken discourse. Appl. Linguist. 24 (4), 425–438.
Baayen, R.H., Piepenbrock, R., Gulikers, L., 1995. The CELEX Lexical Database (Release 2). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.
Balota, D.A., Chumbley, J.L., 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. J. Exp. Psychol. Hum. Percept. Perform. 10, 340–357.
Beglar, D., Nation, P., 2007. A vocabulary size test. Lang. Teach. 31 (1), 9–13.
Bell, H., 2003. Using Frequency Lists to Assess L2 Texts. University of Wales Swansea (Unpublished thesis).
Biemiller, A., Slonim, N., 2001. Estimating root word vocabulary growth in normative and advantaged populations: evidence for a common sequence of vocabulary acquisition. J. Educ. Psychol. 93 (3), 498–520.
Boldt, R.F., Larsen-Freeman, D., Reed, M.S., Courtney, R.G., 1992. Distributions of ACTFL Ratings by TOEFL Score Ranges. TOEFL Res. Rep. 41.
Cobb, T., 1997. Is there any measurable learning from hands-on concordancing? System 25 (3), 301–315.
Cobb, T., 1999. Breadth and depth of lexical acquisition with hands-on concordancing. Comp. Assist. Lang. Learn. 12 (4), 345–360.
Cobb, T., 2000. One size fits all? Francophone learners and English vocabulary tests. Can. Modern Lang. Rev. 57 (2), 295–324.
Cobb, T., 2007. Computing the vocabulary demands of L2 reading. Lang. Learn. Technol. 11 (3), 38–63.
Cobb, T., Horst, M., 1994. Vocabulary sizes of some City University students. City University (HK). J. Lang. Stud. 1 (1), 59–68.
Cobb, T., Horst, M., 2011. Does word coach coach words? CALICO 28 (3), 639–661.
Crossley, S.A., Allen, D., McNamara, D.S., 2011. Text readability and intuitive simplification: A comparison of readability formulas. Read. Foreign Lang. 23 (1), 84–102.
Crossley, S.A., Greenfield, J., McNamara, D.S., 2008. Assessing text readability using cognitively based indices. TESOL Q. 42 (3), 475–493.
Crossley, S.A., Louwerse, M.M., McCarthy, P.M., McNamara, D.S., 2007. A linguistic analysis of simplified and authentic texts. Modern Lang. J. 91 (2), 15–30.
Crossley, S.A., McNamara, D.S., 2012. Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. J. Res. Read. 35 (2), 115–135.
Crossley, S.A., Salsbury, T., 2010. Using lexical indices to predict produced and not produced words in second language learners. Mental Lexicon 5, 115–147.
Crossley, S.A., Salsbury, T., McNamara, D.S., 2010. The role of lexical cohesive devices in triggering negotiations for meaning. Issues Appl. Linguist. 18 (1), 55–80.
Crossley, S.A., Salsbury, T., McNamara, D.S., 2012. Predicting the proficiency level of language learners using lexical indices. Lang. Test. 29 (2), 240–260.
Crossley, S.A., Salsbury, T., McNamara, D.S., Jarvis, S., 2011a. Predicting lexical proficiency in language learners using computational indices. Lang. Test. 28 (4), 561–580.
Crossley, S.A., Salsbury, T., McNamara, D.S., Jarvis, S., 2011b. What is lexical proficiency? Some answers from computational models of speech data. TESOL Q. 45 (1), 182–193.
Davies, A., Widdowson, H., 1974. Reading and writing. In: Allen, J.P., Corder, S.P. (Eds.), Techniques in Applied Linguistics. Oxford University Press, Oxford, pp. 155–201.
Forster, K., Chambers, S., 1973. Lexical access and naming time. J. Verbal Learn. Verbal Behav. 12, 627–635.
Graesser, A.C., Millis, K.K., Zwaan, R.A., 1997. Discourse comprehension. Annu. Rev. Psychol. 48, 163–189.
Graesser, A.C., McNamara, D.S., Louwerse, M., Cai, Z., 2004. Coh-Metrix: Analysis of text on cohesion and language. Behav. Res. Methods Instrum. Comp. 36, 193–202.
Grant, L., Ginther, A., 2000. Using computer-tagged linguistic features to describe L2 writing differences. J. Second Lang. Writ. 9, 123–145.
Horst, M., Collins, L., 2006. From "faible" to strong: how does their vocabulary grow? Can. Mod. Lang. Rev. 63 (1), 83–106.
Hunt, K.W., 1965. Grammatical structures written at three grade levels. NCTE Research Report Number 3 Urbana. National Council of Teachers of English, Illinois.
Jarvis, S., Grant, L., Bikowski, D., Ferris, D., 2003. Exploring multiple profiles of highly rated learner compositions. J. Second Lang. Writ. 12, 377–403.
Kirsner, K., 1994. Implicit processes in second language learning. In: Ellis, N. (Ed.), Implicit and Explicit Learning of Languages. Academic Press, San Diego, CA, pp. 283–312.
Laufer, B., 1992. How much lexis is necessary for reading comprehension? In: Arnaud, P., Béjoint, H. (Eds.), Vocabulary & Applied Linguistics. Macmillan, London, pp. 126–132.
Laufer, B., 1998. The development of passive and active vocabulary in a second language: same or different? Appl. Linguist. 19 (2), 255–271.
Laufer, B., Nation, P., 1995. Vocabulary size & use: lexical richness in L2 written productions. Appl. Linguist. 16 (3), 307–322.
Laufer, B., Nation, P., 1999. A vocabulary size test of controlled productive ability. Lang. Test. 16 (1), 33–51.
Leech, G., Rayson, P., Wilson, A., 2001. Word Frequencies in Written and Spoken English: Based on the British National Corpus. Longman, London.

Lu, X., 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Q. 45 (1), 36–62.
McCarthy, P.M., Jarvis, S., 2010. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behav. Res. Methods 42, 381–392.
McNamara, D.S., Graesser, A.C., 2011. Coh-Metrix: An automated tool for theoretical and applied natural language processing. In: McCarthy, P.M., Boonthum, C. (Eds.), Applied natural language processing and content analysis: Identification, investigation, and resolution. IGI Global, Hershey, PA, pp. 188–205.
Morris, L., Cobb, T., 2004. Vocabulary profiles as predictors of TESL student performance. System 32 (1), 75–87.
Meara, P., 2005. Lexical frequency profiles: a Monte Carlo analysis. Appl. Linguist. 26 (1), 32–47.
Meara, P., Bell, H., 2001. P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. Prospect 16 (3), 5–19.
Meara, P., Buxton, B., 1987. An alternative to multiple choice vocabulary tests. Lang. Test. 4 (2), 142–154.
Murphy, G.L., 2004. The Big Book of Concepts. MIT Press, Cambridge, MA.
Nation, P., 1990. Teaching & Learning Vocabulary. Newbury House, Rowley MA.
Nation, P., 2006. How large a vocabulary is needed for reading and listening? Can. Mod. Lang. Rev. 63 (1), 59–82.
Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W., 2008. Gender differences in language use: an analysis of 14,000 text samples. Discourse Process. 45, 211–246.
Schmitt, N., Schmitt, D., 2013. A reassessment of frequency and vocabulary size in vocabulary teaching. Lang. Teach. (in press).
Ovtcharov, V., Cobb, T., Halter, R., 2006. La richesse lexicale des productions orales: mesure fiable du niveau de compétence langagière. Revue Canadienne des Langues Vivantes 63 (1), 107–125.
Tuldava, J., 1996. The frequency spectrum of text and vocabulary. J. Quant. Linguist. 3 (1), 38–50.
Wendt, A., Woo, A., 2009. A Minimum English Proficiency Standard for the Test of English as a Foreign Language Internet-based Test (TOEFL-iBT). NCLEX Psychometric Research Brief. National Council of State Boards of Nursing.
Zipf, G.K., 1945. The meaning-frequency relationship of words. J. Gen. Psychol. 33, 251–256.

Scott Crossley is an Associate Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, cognitive science, discourse processing, and discourse analysis. His primary research focuses on corpus linguistics and the application of computational tools in second language learning and text comprehensibility.

Tom Cobb is a Professor at the Université du Québec à Montréal. His research interests involve complex knowledge acquisition in second language learners, primarily focused on the lexicon. He has published over 50 peer-reviewed journal articles or book chapters in educational technology, literacy education, or applied linguistics and has developed a multi-plane interactive website which accesses data-driven learning tools and principles for learners, teachers, course developers, and researchers worldwide.

Danielle McNamara is a Professor at Arizona State University and Senior Research Scientist at the Learning Sciences Institute. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research ranges a variety of topics including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms.