He, X., & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, *53*(2), 348-371. https://doi.org/10.1002/tesq.483

Choosing Words to Teach:

A Novel Method for Vocabulary Selection and Its Practical Application

Xuehong (Stella) He and Aline Godfroid

Michigan State University

East Lansing, Michigan, United States

Author Note

Acknowledgements: We thank the instructors who participated in our research. Our special thanks go to Lawrence Zwier, the Associate Director of Curriculum at the English Language Center, Michigan State University, for his insightful feedback on our article.

Corresponding Author: Xuehong (Stella) He, Second Language Studies, Michigan State University, B220 Wells Hall, 619 Red Cedar Rd., East Lansing, MI 48824, USA. Email: hexuehon@msu.edu

Choosing Words to Teach:

A Novel Method for Vocabulary Selection and Its Practical Application

Abstract

Vocabulary learning materials and vocabulary learning research have a common objective of promoting effective vocabulary instruction (Schmitt, 2008), but in practice, vocabulary learning materials tend to reflect materials writers' repertoire and intuition primarily (Tomlinson, 2011). In an effort to develop a stronger interface between research and practice, we introduce a novel method for word selection based on words' frequency, usefulness, and difficulty (Laufer & Nation, 2012). We retrieved the frequency of 191 words and collocations targeted in a North-American IEP from *COCA* and *COCA-Academic*, and collected usefulness and difficulty ratings from 76 experienced ESL instructors. Frequency correlated moderately with usefulness and difficulty, which supported the value of including usefulness and difficulty ratings as word-selection criteria. A cluster analysis revealed five distinct groups of target words, which differed in frequency, usefulness, and difficulty: (i) $Freq_{Hi1}/Use_{Hi1}/Diff_{Lo1}$; (ii) $Freq_{Hi2}/Use_{Hi2}/Diff_{Lo2}$; (iii) $Freq_{Mid}/Use_{Mid}/Diff_{Hi1}$; (iv) $Freq_{Lo2}/Use_{Lo2}/Diff_{Mid}$; (v) $Freq_{Lo1}/Use_{Lo1}/Diff_{Hi2}$. Teaching of the target words could be prioritized according to this sequence. This study introduces a step-by-step approach for materials writers, curriculum designers, and teaching professionals to identify word groupings in a potential list of target words, using a combination of objective and subjective data, with the prospect of creating more effective and more efficacious vocabulary learning materials. (200 words)

Keywords: vocabulary learning, materials development, word selection, frequency, usefulness, difficulty

Choosing Words to Teach:

A Novel Method for Vocabulary Selection and Its Practical Application

Introduction

Learning second language (L2) vocabulary can be a substantial challenge, and success relies on concerted efforts from students, teachers, researchers, and materials writers (Schmitt, 2008). A common view is that researchers probe for evidence of effective vocabulary learning methods, and materials writers deliver research findings to students and teachers in the form of learning materials (Schmitt, 2008). However, the translation of research findings into appropriate learning materials is often wanting, for materials writers may not participate in the evidence-based teaching/learning cycle as fully as expected. Reports have indicated that materials writers often rely heavily on publications, repertoire, and intuition, with little research or even anecdotal support when writing materials (Tomlinson, 2011, 2012a). Commercial publishers of global textbooks are not very likely to adopt research results from L2 acquisition and ironically, published materials are often more closely matched to disputed rather than established theories (Tomlinson, 2012b). A case in point is the presentation of vocabulary in L2 textbooks in semantic sets due to its seeming logic and facilitation of learning (Folse, 2004), even though the weight of the evidence suggests that semantic clustering may hinder learning (Bolger & Zapata, 2011). There is thus a need for a more systematic evaluation and empirical validation of learning materials (Chapelle, 2010; Richards, 2006; Tomlinson, 2012a).

In this study, we evaluated the word selection of the in-house materials created for a dedicated vocabulary course in an Intensive English Program (IEP) in the United States. According to the program website, the IEP offers 20 hours of weekly instruction focusing on academic English. The dedicated vocabulary course was added to the IEP in 2016 in recognition

of the importance of academic vocabulary for students' academic success (Berman & Cheng, 2010; Evans & Green, 2007; Evans & Morrison, 2011; Schmitt, Jiang, & Grabe, 2011; Wu & Hammond, 2011). A team of IEP instructors wrote the in-house materials for the vocabulary course collaboratively. Specifically, to ensure greater continuity within the curriculum, the instructors copied the target words from a textbook used in a different IEP course, namely, *Q: Skills for Success−Reading and Writing 5* (Caplan & Douglas, 2015). The instructors then wrote explicit vocabulary activities for these target words, which served as the in-house materials for the vocabulary course. Copying the word list of the reading textbook allowed for the recycling/repetition of new vocabulary among courses and a better integration of the vocabulary course within the IEP curriculum. However, copying textbook vocabulary was based on the assumption—untested prior to this study—that this textbook would expose students to a range of vocabulary that would be beneficial to their academic studies. In the present study, we set out to test this claim empirically. More generally, the current study addresses a question that will be recognizable to many L2 instructors around the world: with so many words and so little time available, what words should one prioritize for teaching? The method we employed to address this question has wider pedagogical and curricular relevance. It could be implemented in a variety of educational settings. Accordingly, the goals of this study were two-fold: (i) to introduce a novel method for materials writers, curriculum designers, and teaching professionals to evaluate word candidates for teaching empirically, and (ii) to provide empirical evidence to demonstrate this method in the specific context of a North-American IEP.

Word Selection Criteria

A key question in vocabulary teaching is what target words to select for learning (Nation, 2016; Read, 2004). Although recent L2 research has explored a new, personalized approach to

word selection involving student-selected materials (Barker, 2007; Choi & Ma, 2015), selecting words for learning is still a major concern for teachers, researchers, and material writers. Laufer and Nation (2012) proposed the selection criteria for vocabulary be frequency, usefulness, and learnability/difficulty.

Frequency

The cost-benefit principle in vocabulary teaching dictates that learners should get the best return for their learning efforts, and words with high frequency are more likely to provide a better return (Laufer, 2014; Nation, 2011, 2013b; Nation & Webb, 2011). High-frequency words in a general sense refer to words that appear frequently in all kinds of spoken and written texts, regardless of the specific contexts (Nation, 2013a). Without knowledge of and fluent access to these words, learners will suffer in their L2 comprehension and production (Nation, 2013a, 2016). Therefore, words with high frequency ought to be prioritized in teaching and learning, especially for lower-level learners (Laufer, 2014; Nation, 2011, 2013a, 2013b; Nation & Webb, 2011). Lists of high-frequency words, such as the *General Service List of English Words* (West, 1953), have been constructed with this purpose in mind.

Although some specialized words have low frequency counts overall, they can be highly frequent in a certain field and very useful for communication within that field (Nation, 2013a, 2016). The *Academic Vocabulary List* (*AVL*) (Gardner & Davies, 2013) was designed to cater to L2 learners' vocabulary needs in English academic settings. The *AVL* covers nine academic disciplines and is based on *COCA-Academic,* the 120-million-word academic sub-corpus of the 520-million-word *Corpus of Contemporary American English* (*COCA*) (Davies, 2012). In this regard, the *AVL* represents a step forward in terms of corpus size and list creation (Durrant, 2016; Malmström, Pecorari, & Gustafsson, 2016), even though the *AVL*'s ability to represent academic

writing across a range of academic disciplines remains disputed (see Durrant, 2016 and Gardner & Davies, 2016 for discussion).

To estimate word frequency, teachers and materials developers can rely on human intuition or objective corpus frequency counts (Schmitt, 2010). Student intuition can reveal individual experience and exposure to language in the environment (i.e., a second or foreign language learning context) (Schmitt, 2010), whereas teacher intuition may better reflect word frequency in student-directed language input (Wang & Koda, 2005). Regardless of who provides the intuition data, results on the accuracy of human judgment of word frequency have been mixed, with correlations with corpus-based frequency data ranging from .50 (Schmitt & Dunham, 1999) to .70 (Alderson, 2007). On the other hand, corpus analysis supported by technology can provide objective and quantifiable counts of word frequency (Schmitt, 2010). Admittedly, corpora are limited in reflecting individual experience and encompassing private language exchanges (Schmitt, 2010), and methodological problems exist in discerning homonyms and multiword units in corpora (Laufer & Nation, 2012). However, corpora can handle linguistic data that surpass the human brain capacity (Schmitt, 2010), and allow powerful and objectively verifiable analysis of actual language use (Leech, 1992). Actually, most work on word frequency has made use of corpora (Schmitt, 2010).

Usefulness

Although highly robust, frequency should not be the only criterion when considering words for learning (Nation & Webb, 2011). Laufer and Nation (2012) highlighted usefulness as another important criterion for word selection. Since nonnative speakers are likely to master fewer L2 words than native speakers do, vocabulary selected for learning needs to be as useful as possible for functioning in the target language (Laufer & Nation, 2012). Though usefulness and

frequency partially overlap, research has not established that usefulness can be substituted by frequency. Frequent words can be useful for all purposes; however, infrequent words can also be useful in catering to learners' particular needs (Laufer & Nation, 2012). Consequently, corpus data may not reveal the full picture of useful vocabulary. Human intuition also has a role to play.

The role and added value of human intuition is apparent in the construction of lists of multiword units, for which usefulness is a desirable feature. The last decade has witnessed the publication of several such lists, including the *Academic Formulas List* (Simpson-Vlach & Ellis, 2010), the *Phrasal Expressions List* (Martinez & Schmitt, 2012), the *Academic Collocation List* (Ackermann & Chen, 2013), and the *Phrasal Verb Pedagogical List* (Garnier & Schmitt, 2015). In compiling these lists, researchers either used human intuition as a direct criterion for word inclusion or exclusion (Ackermann & Chen, 2013; Garnier & Schmitt, 2015; Martinez & Schmitt, 2012) or used human intuition as a dependent variable in a statistical analysis to weight different corpus-derived metrics statistically (Simon-Vlach & Ellis, 2010). The common idea behind all of these studies is that word selection that combines subjective intuition and objective frequency counts can yield more meaningful and appealing lists for teaching and learning (Simpson-Vlach & Ellis, 2010). As Martinez and Schmitt (2012) noted, human intuition can provide key qualitative judgment when identifying and selecting multiword units for teaching, which is not likely to be replicated by contemporary computer programs. Given the versatility and flexibility of human intuition, we collected teacher ratings in this study to elicit information about the perceived usefulness and learnability/difficulty of the target words.

Learnability/Difficulty

The third criterion of word selection proposed by Laufer and Nation (2012) is the learnability, or difficulty, of a word. Word difficulty can be divided into two types, interlingual,

which results from the interactions between the first language (L1) and the L2, and intralingual, which comes from the interactions between new words and familiar words in the target language (Laufer, 2014; Laufer & Nation, 2012; for reviews of word difficulty, see Laufer, 1990b, 1997). Word difficulty is affected by the pronounceability, regularity of spelling, and part of speech of a word, as well as word length and number of syllables, morphological transparency, and concreteness of meaning and imageability (Ellis & Beaton, 1993; Nation & Webb, 2011). Generally speaking, both easy and difficult words are worth studying. Learning easy words can enhance communicative abilities without requiring much learning effort, while teaching difficult words can help reduce language errors associated with difficult vocabulary (Laufer, 1990a). Nevertheless, to maximize the learning return from the limited teaching time, difficult words deserve more time for explanation and practice in class (Laufer, 1990a, 1990b, 1997), as the easy vocabulary could be learned by the students after class without much effort.

Although difficult words are often defined as low-frequency words, a frequency-based definition is insufficient because it leaves out other important factors such as context of use (Meara & Bell, 2001). Experienced teachers, on the other hand, can make informative judgment of word properties beyond frequency, by utilizing their knowledge of textbooks and materials and considering additional factors such as cognateness (Tidball & Treffers-Daller, 2008). As a case in point, Bardel, Gudmundson, and Lindqvist (2012) recently employed teacher judgment of word difficulty to develop a modified profiling method of learners' oral production. The researchers collected word difficulty ratings by asking teachers (both native and nonnative speakers of the target language) to indicate on a six-point scale whether a word was basic or advanced. Results showed that the modified method produced better results (i.e., larger effect sizes) than a profiling method that was purely frequency-based. Specifically, the modified

method differentiated better among Italian and French learners at different proficiency levels, and thus provided more accurate lexical profiles of L2 learners.

Other studies on lexical richness and lexical sophistication of learner production have also operationalized word difficulty as whether teachers judge the word as basic or advanced. For example, Daller, van Hout, and Treffers-Daller (2003) asked teachers to indicate whether some Turkish words were basic or more advanced based on their teaching experience. Results showed that teacher judgment was highly reliable, with .92 inter-rater reliability and a within-rater reliability between .86 and .87. Given the high reliability, Daller et al. decided to use teacher ratings instead of a frequency-based vocabulary list by Tezcan (1988) to classify words in an oral production task. In another study comparing three different ways of operationalizing lexical sophistication, Tidball and Treffers-Daller (2008) found that native and bilingual teachers' judgment was a more reliable index than other methods, including frequency of lexical items. A third study conducted by Witalisz and Lesniewska (2007) compared student perception and teacher perception of word difficulty in a student-written essay. Results showed that students and teachers shared similar perception of word difficulty, although teachers' academic backgrounds could affect their judgment.

Aims of the Study

Drawing on the widely accepted criteria of frequency, usefulness, and difficulty, we propose a novel method to evaluate lists of potential target words for inclusion in a textbook or lexical syllabus. We illustrate the method with the case of a dedicated vocabulary course in an American IEP, whose target word selection was based on the reading textbook (Caplan & Douglas, 2015) used in the same program. The following research questions (RQs) guided the study:

1 What is the frequency, usefulness, and difficulty profile of the target words in the vocabulary materials of an upper-intermediate pre-university ESL course?

2 Can the target words in the vocabulary materials be assigned to different groupings to prioritize for teaching based on the words' frequency, usefulness, and difficulty?

Methods

Participants

Seventy-six ESL instructors participated in this study as expert raters. Sixty-eight instructors were English native speakers and eight were nonnative speakers, whose native languages were French, Italian, Nepali, Polish, Russian, Sinhala, and Turkish. Two instructors had received doctoral degrees and the remaining 74 instructors held master's degrees. On average, the instructors had taught English for a total of 11.72 years (*SD* = 8.92), 9.81 years (*SD* = 7.06) of which were at the college level in North America.

Materials

Corpora for Frequency Statistics. To obtain reliable frequency counts, we looked up the target words in *COCA* and *COCA-Academic* (Davies, 2012). *COCA* currently contains over 520 million words of American English texts from 1990 to 2015, and its sub-corpus *COCA-Academic* includes 120 million words of academic texts from 1990 to 2011. *COCA-Academic* was the basis for compiling the *AVL*, which contains the most frequent 20,845 lemmas, and the *Word Families List* (*AVL-Families*), which contains the 1,991 core word families in contemporary American English (Gardner & Davies, 2013). Because the vocabulary course was geared towards students preparing for academic studies in the United States, we adopted *COCA-Academic* as the major corpus and *COCA* as its supplement.

The vocabulary materials for this study initially comprised 191 unique target words. Except for two words, *olive-toned* and *sun-kissed,* the target words all appeared in *COCA-Academic*. To allow for a meaningful comparison of word frequency data, *olive-toned* and *sun-kissed* were excluded from further analysis. Thus, the final word set consisted of 189 words: 165 single words and 24 multiword units.

To retrieve the raw frequency of a single word, we first looked up the *AVL-Families* and recorded the raw frequency of the exact word instead of the word family, taking part of speech into consideration. If the single word was not in the *AVL-Families*, we looked it up in the *AVL*. If the word was still not there, we checked *COCA-Academic* and recorded the raw frequency. For the raw frequency of a multiword unit, we looked up all inflected forms of the multiword unit in *COCA-Academic* and then added up the raw frequencies. For instance, for the multiword unit *take issue with*, we searched and added up the raw frequencies of *take issue with*, *takes issue with*, *taking issue with*, *took issue with*, and *taken issue with*. For each single word or multiword unit, we calculated normalized frequency (frequency per million words) by dividing the raw frequencies by 120.032441 (total million words in *COCA-Academic*).

Online Survey for Usefulness and Difficulty Ratings. An online survey was created to collect ESL teachers' ratings of word usefulness and difficulty. The survey consisted of two sections. The first section asked for teachers' background, including their L1, academic background, and teaching experience. The other section asked for teachers' ratings of the target words' perceived usefulness and difficulty. This section contained 67 words accompanied by their part of speech, with two rating questions for every word. Teachers first rated the usefulness of all words ("*Whether the vocabulary is worth teaching students who are preparing to go to college in North America*") on a 7-point scale, with 1 for "*Not worth*" and 7 for "*Indispensable*",

adapted from Simpson-Vlach and Ellis (2010). They then rated each word's difficulty ("*How advanced is the vocabulary*") on a 7-point scale, with 1 for "*Basic/Easy*" and 7 for "*Advanced/Difficult*", following Bardel et al.'s (2012) labelling. Among the 67 words in the survey, 42 were anchor words, meaning that they stayed the same across six versions of the survey. The remaining 25 words were unique to each version. Therefore, each of the 42 anchor words was rated by at least 73 teachers, and each of the remaining 149 words was rated by at least 11 teachers. Cronbach's alpha for the anchor words was .947 for usefulness and .973 for difficulty, indicating that teachers' ratings were highly reliable. Consequently, we averaged teachers' usefulness ratings and difficulty ratings, which yielded one usefulness and one difficulty score per word.

Procedure

To recruit teachers, we searched the IEP websites of major universities across the United States for ESL instructors' contact information. We sent out individual emails to ESL instructors, in which we briefly introduced our project, including its significance for IEP teaching, and invited the instructors to complete an online survey in *Qualtrics*. The survey asked instructors to rate every word in terms of its usefulness and then, in a second round after that, its difficulty. The whole procedure took about 10 minutes to complete. Participants received a US$5 gift card for their time.

Data Analysis

To answer RQ1, we first used bootstrapping to calculate the descriptive statistics for the target words' frequency, usefulness, and difficulty data. Bootstrapping is a type of robust statistics that uses the current sample as the population and draws new samples from it, on which the same statistical test is performed repeatedly. Bootstrapping has the potential to overcome

problems associated with small sample sizes and non-normal distributions and, because of this, it can provide robust results and more statistical power (LaFlair, Egbert, & Plonsky, 2015; Larson-Hall, 2016; Larson-Hall & Herrington, 2009). Next, we performed bivariate correlation analyses on the frequency, usefulness, and difficulty data for the vocabulary items. Before performing the correlations, we checked assumptions of independence of observations, linearity, normality, and homoscedasticity (see Larson-Hall, 2016). Independence of observations was fulfilled with the current research design, and frequency, usefulness, and difficulty data were confirmed to be in a linear relationship. However, visual inspection and statistical testing indicated that the data were not normally distributed (Shapiro-Wilk test, frequency, $W(189) = .622$, $p < .001$; usefulness, $W(189) = .957$, $p < .001$; difficulty, $W(189) = .985$, $p = .041$) and the frequency data had unequal variance (Levene's test, $F(1, 187) = 3.957$, $p = .048$). To accommodate these violations of assumptions, we opted for bootstrapped Pearson's correlations for the analysis.

For RQ2, we performed a cluster analysis on the frequency, usefulness, and difficulty data to identify word groups that patterned together along these dimensions (see Staples & Biber, 2015). By introducing cluster analysis to the problem of target word selection, we were able to compare the potential learning return of different word groups, and to provide a basis for prioritizing their teaching. One problem to identify and address in cluster analysis is collinearity, which means clustering variables are highly correlated (Ketchen & Shook, 1996; Mooi & Sarstedt, 2011; Sambandam, 2003). Correlations higher than .50 (Sambandam, 2003) or .90 (Mooi & Sarstedt, 2011) are regarded as signs of collinearity in cluster analysis, and in multivariate statistics more generally, $r > .70$ is often proposed as a cutoff value for collinearity (e.g., Dormann et al., 2013). Results from RQ1 showed that except for the correlation between frequency and usefulness for multiword units ($r = .556$), correlations were lower than the

strictest .50 benchmark (see Table 2). Thus we did not regard collinearity as a serious problem for performing cluster analysis with our data. Since cluster analysis makes no assumptions about data distribution, we opted for a hierarchical cluster analysis for the relatively small sample (Norušis, 2011). Hierarchical cluster analysis is the dominant approach to cluster analysis in L2 research and does not require a predetermined number of clusters (Staples & Biber, 2015). In our analysis, the clustering variables were normalized frequency and usefulness and difficulty ratings, the clustering method was Ward's method, and the distance measure was the squared Euclidean distance. We standardized all data by converting them into *z* scores, which removed the potential impact of having different scales in the variables (see Csizér & Jamieson, 2012).

Results

RQ1: Frequency, Usefulness, Difficulty, and Their Correlations

The complete list of frequency, usefulness, and difficulty data for the 191 target words is available in *Supplementary Materials A*. Table 1 shows the bootstrapped descriptive statistics of the 189 target words that appeared in *COCA-Academic*, which were obtained from 10,000 bootstrapping samples (LaFlair et al., 2015; Larson-Hall, 2016). BCa (bias corrected and accelerated) 95% *CI*s were reported for the means because the BCa method is generally more accurate (LaFlair et al., 2015). The average word frequency was 31.02 per million (*SD* = 46.20) and usefulness and difficulty ratings were 5.31 (*SD* = 0.99) and 4.35 (*SD* = 0.77) out of 7, respectively. A word frequency of 31.02 falls within the most frequent 3,000 lemmas of the *AVL*, meaning these target words were generally frequent academic words. Usefulness and difficulty ratings showed that on average, the target words were regarded as useful and of medium difficulty. Results of single words and multiword units revealed that they were perceived as similar in terms of difficulty, and that single words were regarded as somewhat more useful.

Whereas on average, single words belonged to the most frequent 3,000 lemmas, multiword units fell within the most frequent 6,000 lemmas.

[Insert Table 1 Here]

Table 2 shows the results of bootstrapped Pearson's correlations of the three criteria. For the 189 target words that appeared in *COCA-Academic*, the correlations between frequency and usefulness, between frequency and difficulty, and between usefulness and difficulty were statistically significant with about medium effect sizes, according to Plonsky and Oswald's (2014) recommended benchmark of $r = .40$ for a medium effect size. The results for the 165 single words mirrored those of the entire data set; however, for the 24 multiword units, the correlation between frequency and usefulness rose to .56. Difficulty did not correlate with either usefulness or frequency in the case of multiword units. With the exception of these two correlations, the BCa 95% *CI*s of the correlation coefficient *r*s did not span 0, which confirmed the statistical significance of the results.

[Insert Table 2 Here]

RQ2: Vocabulary Clustering Based on Frequency, Usefulness, and Difficulty

The goal of a cluster analysis is to identify patterns of data within a multidimensional space. In the case of this study, we aimed to model words (single words and multiword units) situated in a three-dimensional space, defined by frequency, usefulness, and difficulty. After the optimal number of clusters has been decided (see next paragraph), we will describe the resulting clusters in terms of their frequency (*Freq*.), usefulness (*Use*.), and difficulty (*Diff*.). To do so, we will distinguish high (*Hi*), medium (*Mid*) and low (*Lo*) levels of each word property. For instance, $Freq_{Hi}/Use_{Hi}/Diff_{Lo}$ denotes the set of words that is highly frequent, highly useful, and low in difficulty (i.e., a high priority for teaching). The *Hi/Mid/Lo* descriptors are relative to other target

words in the vocabulary materials. In some cases, we will index the levels further, for instance, $Freq_{Hi1}$ (the highest frequency) or $Freq_{Hi2}$ (the second highest frequency), when more fine-grained distinctions are needed.

To decide the optimal number of clusters in the hierarchical cluster analysis, we inspected the coefficients of the agglomeration schedule in the statistical output (see Staples & Biber, 2015). When the difference between the coefficients of two adjacent clustering stages begins to flatten out, little new information is added by further dividing the existing clusters, and the optimal number of clusters is likely to have been reached (Staples & Biber, 2015). *Supplementary Materials B* shows the coefficients and their differences when the final 10 clusters were formed. Figure 1 plots the differences in coefficients against the numbers of clusters to allow for a visual inspection of the flattening point. The figure shows that starting from four clusters, the slope decreased and that it flattened out obviously at six clusters. Therefore, the optimal number of clusters was likely to range between four and six.

[Insert Figure 1 Here]

To further determine the optimal number of clusters, we compared the three clustering solutions (i.e., four, five, or six clusters). The complete list of the cluster membership of every word in each clustering solution is attached in *Supplementary Materials C*. Table 3 presents the bootstrapped descriptive statistics of the three clustering solutions. To understand the distinctive characteristics of each cluster group, we compared the means of frequency, usefulness, and difficulty of different cluster groups (see Staples & Biber, 2015). Due to the small sample size of *Cluster 6* (n = 3) in the *Six-Cluster Solution* (see Table 3), we decided not to perform any statistical analysis on this solution. Instead, we conducted bootstrapped one-way ANOVAs on the *Four-* and *Five-Cluster Solutions* (bootstrapping to account for deviations from normality

and homogeneity of variance).[1] Results showed that in the *Four-* and *Five-Cluster Solutions*, the cluster groups differed significantly on each criterion variable (frequency, usefulness, and difficulty) at $p < .001$. This supports the notion that the words in different clusters were distinct sets that may be prioritized for teaching differently. At the same time, the resulting clusters in a cluster analysis will almost always have statistically significant differences among them (Csizér & Jamieson, 2012), because this is what cluster analysis does—it maximizes group differences. Therefore, these results should be used for confirmatory purposes only.

[Insert Table 3 Here]

Next, we generated a 3D scatterplot for the *Six-Cluster Solution* with frequency (per million words), usefulness, and difficulty data plotted on the three dimensions of the graph (see Figure 2). As shown in Figure 2, *Clusters 5* and *6* were in adjacent regions; they had the highest mean frequency and usefulness and the lowest mean difficulty (see Table 3). Therefore, we combined *Clusters 5* and *6* into one group. What remained then was to choose between the *Five-* and *Four-Cluster Solutions*. These solutions differed with regard to the composition of the first cluster as either a large set of 58 words in the *Four-Cluster Solution* or two smaller clusters of 35 and 23 words, respectively, in the *Five-Cluster Solution* (see Table 3).

[Insert Figure 2 Here]

The data in Table 3 confirmed that *Clusters 1* and *2* in the *Five-Cluster Solution* were comparable in frequency, but differed widely in usefulness and difficulty. This supported keeping *Clusters 1* and *2* as separate groups (also see Figure 2). In practical terms, this meant that words similar in frequency but different in usefulness and difficulty should be treated differently to increase potential learning return. Therefore, the optimal clustering solution was to divide the 189 target words into five groups.

Table 4 summarizes the word characteristics of the five clusters, which represent different levels of teaching priority. *Cluster 5* had the highest frequency and usefulness and the lowest difficulty. It was *Group* $Freq_{Hi1}/Use_{Hi1}/Diff_{Lo1}$. On average, words in this group fell within the most frequent 1,000 lemmas of the *AVL*. An example was *consequence*, which had high frequency (115.52) and usefulness (6.92), and low difficulty (3.50). *Cluster 4* had the second highest frequency and usefulness and the second lowest difficulty. It was *Group* $Freq_{Hi2}/Use_{Hi2}/Diff_{Lo2}$. Words in this group fell within the most frequent 3,000 lemmas of the *AVL* on average. An example was *ethnicity*, which had high frequency (33.02) and usefulness (6.01), and low difficulty (3.92).

[Insert Table 4 Here]

Next was *Cluster 3*, with medium frequency and usefulness but the highest difficulty. This cluster was *Group* $Freq_{Mid}/Use_{Mid}/Diff_{Hi1}$. On average, words in this group fell within the most frequent 4,000 lemmas of the *AVL*. An example is *erosion*, which had medium frequency (21.23) and usefulness (4.67), and high difficulty (5.46). Finally, two clusters combined low frequency and usefulness with medium or high difficulty. *Cluster 1* had the second lowest frequency and usefulness and medium difficulty. It was *Group* $Freq_{Lo2}/Use_{Lo2}/Diff_{Mid}$. The average frequency of words in this group corresponded to the most frequent 7,000 lemmas of the *AVL*. One example was *skyrocket*, which had low frequency (1.92) and usefulness (4.00), and medium difficulty (4.33). *Cluster 2* had the lowest frequency and usefulness but the second highest difficulty. It was *Group* $Freq_{Lo1}/Use_{Lo1}/Diff_{Hi2}$. On average, words in this group fell beyond the most frequent 10,000 lemmas of the *AVL*. An example was *incinerate*, which had low frequency (0.98) and usefulness (3.15), and high difficulty (4.83).

Discussion

Taking the target words in real-world vocabulary materials as an example, we aimed to introduce a novel method for evaluating word selection in ESL/EFL settings. Word selection is the foundation of any type of vocabulary instruction. Therefore, the issue of what words to teach deserves careful consideration in both research and practice (Nation, 2013a, 2013b, 2016; Read, 2004). Through the empirical analysis of an actual word list, we developed a protocol for choosing target words that could be adopted in a variety of educational settings. Our discussion will reflect these two goals. We will consider, first, how target words in the current vocabulary materials can be prioritized for teaching and, second, how materials writers and curriculum designers can proceed to select target words in the context of their own educational projects.

Prioritizing Target Words in the Vocabulary Materials

To identify larger groupings in a word list, we performed a cluster analysis on the lexical selection criteria of the 189 target words that appeared in *COCA-Academic*. These words were clustered into five groups: (i) $Freq_{Hi1}/Use_{Hi1}/Diff_{Lo1}$; (ii) $Freq_{Hi2}/Use_{Hi2}/Diff_{Lo2}$; (iii) $Freq_{Mid}/Use_{Mid}/Diff_{Hi1}$; (iv) $Freq_{Lo2}/Use_{Lo2}/Diff_{Mid}$; (v) $Freq_{Lo1}/Use_{Lo1}/Diff_{Hi2}$. The clustering solution showed that frequency and usefulness patterned together (e.g., both high or both low), whereas difficulty was often, but not always, inversely related to the other criterion variables (compare *Clusters 4* and *5*). Although this result was intuitive, it was perhaps unexpected considering frequency–usefulness–difficulty correlations for single words were only small to medium, in the .30 - .40 range (see Table 2).

Given the cost-benefit principle in word selection (Laufer, 2014; Nation, 2011, 2013b; Nation & Webb, 2011), highly frequent and useful words with low difficulty are the most likely to provide a good return for learners' efforts. This is because learners do not need to expend much mental energy to learn these words but can improve considerably in their communicative

abilities. Thus, *Group* $Freq_{Hi1}/Use_{Hi1}/Diff_{Lo1}$ deserves the first priority in teaching, followed by *Group* $Freq_{Hi2}/Use_{Hi2}/Diff_{Lo2}$. Another consideration is that, given the limited time for teaching, more explanation and practice should be devoted to difficult words, so as to assist learners in their progress (Laufer, 1990a, 1990b, 1997). In this view, *Group* $Freq_{Mid}/Use_{Mid}/Diff_{Hi1}$ should be introduced after *Group* $Freq_{Hi2}/Use_{Hi2}/Diff_{Lo2}$, and before *Group* $Freq_{Lo2}/Use_{Lo2}/Diff_{Mid}$. The lowest priority should be *Group* $Freq_{Lo1}/Use_{Lo1}/Diff_{Hi2}$. In an effort to optimize teaching, words in this group could be replaced by more frequent and useful academic vocabulary.

The groups identified here represent specific sets of words in the vocabulary materials (see *Supplementary Materials C*). With different word lists or courses at different proficiency levels, the actual composition of each cluster will vary, but similar clustering patterns might still obtain. The present findings, therefore, (i) highlight the value of considering multiple indices in word selection using multivariate statistics, (ii) support a role for human judgment in addition to corpus-based frequency data, and (iii) suggest difficulty ratings in particular may offer new information relative to frequency data.

Selecting Target Words for Teaching

Word selection in this study was guided by the frequency, usefulness, and difficulty of the lexical items (Laufer & Nation, 2012). Correlation analyses revealed that frequency correlated moderately with the other two criteria for single words. For multiword units, only frequency and usefulness correlated significantly, whereas the correlation between frequency and difficulty was approaching zero. This supports the idea that frequency, usefulness, and difficulty provide complementary criteria for word selection (Laufer & Nation, 2012) and that different methods may be necessary for measuring different indices.

The current study provides a specific example of how such a multivariate approach could work. Specifically, a protocol for word selection may consist of the following steps: (1) select high-frequency words from existing word lists, corpora, or other teaching materials, (2) choose a representative corpus and retrieve the frequency count for each word, (3) have these words rated by teachers for their usefulness and difficulty, (4) conduct a cluster analyses with frequency, usefulness, and difficulty data or, as a simpler alternative, inspect the data visually, and (5) identify word groupings to prioritize for teaching. It is important to first sort out words with high frequency and usefulness, and then to plan the teaching according to these words' difficulty levels, with more time devoted to explaining and practicing difficult words (Laufer, 1990a, 1990b, 1997).

In practice, retrieving high-frequency words from existing word lists or corpora is a viable option for materials developers and teachers alike. Many word lists and corpora are available online, including the *Academic Word List (AWL)* (Coxhead, 2000), *AVL* (Gardner & Davies, 2013), *General Service List of English Words* (West, 1953), *COCA* (Davies, 2012), and *British National Corpus*. Recently, lists of multiword units have also been compiled, such as the *Academic Collocation List* (Ackermann & Chen, 2013), *Academic Formulas List* (Simpson-Vlach & Ellis, 2010), *Phrasal Expressions List* (Martinez & Schmitt, 2012), and *Phrasal Verb Pedagogical List* (Garnier & Schmitt, 2015). Another approach, which was pursued in the IEP with whom we collaborated, was to adopt the word list from a published textbook. Compared to word lists in peer-reviewed research articles, textbook writers may be less transparent about the principles that guided their word selection process. In such cases, empirical validation of vocabulary materials is especially warranted to ensure that the benefits from recycling words across multiple courses are truly worth students' and teachers' time.

Although frequency data are now widely available and accessible to conduct Steps (1) and (2) of the protocol, time constraints or a lack of resources may make it difficult for teachers to collect usefulness and difficulty ratings from a representative sample (Step [3]). Furthermore, the technical expertise required for Step (4), the cluster analysis, may be prohibitive for many individual teachers. We envision that the proposed method will be of interest to many TESOL professionals, but in many cases, implementation will fall on curriculum designers, materials writers, and publishing companies. In our opinion, these three parties are indispensable players in vocabulary instruction who can play a pivotal role in promoting evidence-based teaching practices. The method we have presented in this study is not difficult to implement for people trained in statistics—in fact, it would make a good project for a graduate-level course on research methodology. Compared to other costs that go into developing teaching materials, this method is therefore inexpensive. Curriculum designers and materials writers could negotiate funding with their program directors or publishing companies to hire freelance statistical consultants. Published textbooks have a major impact on curriculum design and course planning. Therefore, the prospect of being able to support word selection in teaching materials empirically should entice program administrators and textbook editors to support this initiative.

This raises the larger question of what types of projects this method could profitably be employed for, besides the validation of existing materials. Larger programs could plan a curricular project in which they evaluate the vocabulary across their different courses using the present protocol. The results of the vocabulary evaluation could be used to guide vocabulary teaching in both traditional, skill-based courses and dedicated vocabulary courses. For example, word lists could be trimmed or adapted, thereby giving students the opportunity to focus more on important vocabulary in their courses. In a reading course, the teacher could use the prioritized

word groupings to balance class time between practicing reading skills and expanding students' vocabulary. Another scenario, which motivated the current study, is the development of a new vocabulary course to strengthen and expand students' lexical repertoires. Although the previously mentioned word lists offer a good starting point, many of them would require trimming for use in the classroom. Cluster analysis could be a tool to do this. Trimming word lists may also be desirable when a course is shortened or moved to a different medium (e.g., online). In short, there are many scenarios where an empirical approach to word selection can provide benefits.

Limitations and Future Research

One limitation of our study is related to the bootstrapping method in our data analyses. An assumption of bootstrapping is that the sample is sufficiently representative (LaFlair et al., 2015; Larson-Hall & Herrington, 2009); however, the respondents to our online survey were a self-selected and probably highly motivated group of teachers who volunteered their time for a small compensation. Thus, sample representativeness is not guaranteed. Another limitation is that our current groupings of words were based on a specific textbook for a particular proficiency level. Future researchers can evaluate whether the observed patterning of word-level variables also holds in other contexts and thus may reflect some general properties of the English language. Finally, item-level statistics revealed that multiword units generally had much lower frequency than single words did (but had comparable levels of usefulness and difficulty). In light of these inherent differences, it may be better to evaluate the frequency of a lexical item relative to other members in its category and to analyze single words and multiword units separately. Despite these limitations, we believe the method proposed in this study allows researchers to further refine established, mainly frequency-driven word lists, such as the *AWL* (Coxhead, 2000) and the

AVL (Gardner & Davies, 2013). The long-range goal, then, is to contribute to ongoing initiatives to make vocabulary instruction more effective.

Conclusion

In line with the call for research-based evaluation of materials development (Tomlinson, 2012a), we showed how frequency, usefulness, and difficulty information can be combined for the purpose of evaluating word selection empirically. Based on our findings, we also proposed a protocol that materials writers, curriculum designers, and other interested individuals can follow to select words to teach. Although the exact word groupings observed in this study may not apply to a different textbook and/or a different proficiency level, the protocol for word selection itself can be extended to other educational contexts and student populations. We invite other researchers and TESOL professionals to join this effort and evaluate the words that populate their word lists and teaching materials, for a more transparent and more principled approach to vocabulary instruction.

References

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383–409. https://dx.doi.org/10.1093/applin/amm024

Bardel, C., Gudmundson, A., & Lindqvist, C. (2012). Aspects of lexical sophistication in advanced learners' oral production. *Studies in Second Language Acquisition*, *34*(2), 269–290. https://dx.doi.org/10.1017/S0272263112000058

Barker, D. (2007). A personalized approach to analyzing "cost" and "benefit" in vocabulary selection. *System*, *35*(4), 523–533. https://dx.doi.org/10.1016/j.system.2007.09.001

Berman, R., & Cheng, L. (2010). English academic language skills: Perceived difficulties by undergraduate and graduate students, and their academic achievement. *Canadian Journal of Applied Linguistics (CJAL)/Revue Canadienne de Linguistique Appliquée (RCLA)*, *4*(1), 25–40.

Bolger, P., & Zapata, G. (2011). Semantic categories and context in L2 vocabulary learning. *Language Learning*, *61*(2), 614-646. https://dx.doi.org/10.1111/j.1467-9922.2010.00624.x

Caplan, N. A., & Douglas, S. R. (2015). *Q: Skills for success reading and writing 5*. New York: Oxford University Press.

Chapelle, C. A. (2010). The spread of computer-assisted language learning. *Language Teaching*, *43*(1), 66-74. https://dx.doi.org/10.1017/S0261444809005850

Choi, M. L., & Ma, Q. (2015). Realising personalised vocabulary learning in the Hong Kong context via a personalised curriculum featuring "student-selected vocabulary". *Language and Education*, *29*(1), 62–78. https://dx.doi.org/10.1080/09500782.2014.942318

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*(2), 213–238. https://dx.doi.org/10.2307/3587951

Csizér, K., & Jamieson, J. (2012). Cluster analysis. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Chichester, West Sussex, UK: Wiley-Blackwell. https://dx.doi.org/10.1002/9781405198431.wbeal0138

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222. https://dx.doi.org/10.1093/applin/24.2.197

Davies, M. (2012). Corpus of Contemporary American English (1990–2012). Retrieved from http://corpus.byu.edu/coca/

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes*, *43*, 49–61. https://10.1016/j.esp.2016.01.004

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, *43*(4), 559-617. https://dx.doi.org/10.1111/j.1467-1770.1993.tb00627.x

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes*, *6*(1), 3–17. http://doi.org/10.1016/j.jeap.2006.11.005

Evans, S., & Morrison, B. (2011). The first term at university: Implications for EAP. *ELT Journal*, *65*(4), 387–397. http://doi.org/10.1093/elt/ccq072

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor: The University of Michigan Press. https://dx.doi.org/10.3998/mpub.23925

Gardner, D., & Davies, M. (2013). A new Academic Vocabulary List. *Applied Linguistics*, *35*(3), 305–327. https://dx.doi.org/10.1093/applin/amt015

Gardner, D., & Davies, M. (2016). A response to "To what extent is the Academic Vocabulary List relevant to university student writing?" *English for Specific Purposes*, *43*, 62-68. https://doi.org/10.1016/j.esp.2016.03.001

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645–666. https://dx.doi.org/10.1177/1362168814559798

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, *17*(6), 441-458.

LaFlair, G., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46–77). New York; London: Routledge.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. London; New York: Routledge.

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(3), 368-390.

Laufer, B. (1990a). Ease and difficulty in vocabulary learning: Some teaching implications. *Foreign Language Annals*, *23*(2), 147–155. https://dx.doi.org/10.1111/j.1944-9720.1990.tb00355.x

Laufer, B. (1990b). Words you know: How they affect the words you learn. In J. Fisiak (Ed.), *Further insights into contrastive linguistics* (pp. 573–593). Amsterdam: John Benjamins. https://dx.doi.org/10.1075/llsee.30.35lau

Laufer, B. (1997). What's in a word that makes it hard or easy? Intralexical factors affecting vocabulary acquisition. In N. Schmitt & M. J. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 140–155). Cambridge: Cambridge University Press.

Laufer, B. (2014). Vocabulary in a second language: Selection, acquisition, and testing: A commentary on four studies for JALT Vocabulary SIG. *Vocabulary Learning and Instruction*, *3*(2), 38–46.

Laufer, B., & Nation, I. S. P. (2012). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). London; New York, NY: Routledge. https://dx.doi.org/10.4324/9780203808184.ch10

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Startvik (Ed.), *Directions in corpus linguistics* (pp. 105–122). Berlin: Mouton de Gruyter. https://dx.doi.org/10.1515/9783110867275.105

Malmström, H., Pecorari, D., & Gustafsson, M. (2016). Coverage and development of academic vocabulary in assessment texts in English Medium Instruction. In S. Göpferich & I. Neumann (Eds.), *Assessing and developing academic and professional writing skills* (pp. 45-69). New York: Peter Lang.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics*, *33*(3), 299–320. https://dx.doi.org/10.1093/applin/ams010

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.

Mooi, E., & Sarstedt, M. (2011). *A concise guide to market research.* Berlin: Springer-Verlag.

Nation, I. S. P. (2011). Research into practice: Vocabulary. *Language Teaching*, *44*(4), 529–539. https://dx.doi.org/10.1017/S0261444811000267

Nation, I. S. P. (2013a). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2013b). Materials for teaching vocabulary. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 351–365). London: Bloomsbury Academic.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage Learning.

Norušis, M. (2011). Cluster analysis. In *IBM SPSS Statistics 19 statistical procedures companion* (pp. 375–404). Upper Saddle River, N.J.: Prentice Hall.

Plonsky, L., & Oswald, F. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://dx.doi.org/10.1111/lang.12079

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics*, *24*, 146–161. https://dx.doi.org/10.1017/S0267190504000078

Richards, J. (2006). Materials development and research–making the connection. *Regional Language Centre Journal*, *37*(1), 5-26. https://dx.doi.org/10.1177/0033688206063470

Sambandam, R. (2003). Cluster analysis gets complicated. *Marketing Research*, *15*(1), 17-21.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York: Palgrave Macmillan. https://dx.doi.org/10.1057/9780230293977

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research*, *15*, 389–411. https://dx.doi.org/10.1191/026765899669633186

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43. https://dx.doi.org/10.1111/j.1540-4781.2011.01146.x

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512. https://dx.doi.org/10.1093/applin/amp058

Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 243–274). New York; London: Routledge.

Tezcan, N. (1988). *Elementarwortschatz Tukisch-Deutsch*. Wiesbaden: Otto Harrassowitz.

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgments can tell us about basic and advanced words. *French Language Studies*, *18*(3), 299–313. https://dx.doi.org/10.1017/S0959269508003463

Tomlinson, B. (2011). Introduction: Principles and procedures of materials development. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 1–34). New York: Cambridge University Press.

Tomlinson, B. (2012a). Materials development for language learning and teaching. *Language Teaching*, *45*(2), 143–179. https://dx.doi.org/10.1017/S0261444811000528

Tomlinson, B. (2012b). Second language acquisition and materials development. In B. Tomlinson (Ed.), *Applied linguistics and materials development* (pp. 11–30). London; New York, NY: Bloomsbury Academic.

Wang, M., & Koda, K. (2005). Commonalities and differences in word identification skills among learners of English as a second language. *Language Learning*, *55*(1), 71–98. https://dx.doi.org/10.1111/j.0023-8333.2005.00290.x

West, M. (1953). *A General Service List of English words*. London: Longman, Green and Co.

Witalisz, E., & Lesniewska, J. (2007). Perception of word difficulty in L2: Teacher vs. learner judgments. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, *124*, 193–201.

Wu, W., & Hammond, M. (2011). Challenges of university adjustment in the UK: A study of East Asian master's degree students. *Journal of Further and Higher Education*, *35*(3), 423–438. http://doi.org/10.1080/0309877X.2011.569016

Table 1

Bootstrapped Descriptive Statistics of 189 Target Words

			Frequency	Usefulness	Difficulty
All Words	*Mean*		31.02	5.31	4.35
n = 189	*BCa 95%CI*	*Lower*	25.15	5.16	4.24
	of Mean	*Upper*	37.62	5.45	4.46
	SD		46.20	0.99	0.77
	Min		0.02	2.75	2.58
	Max		338.98	7.00	6.08
Single Words	*Mean*		34.26	5.44	4.34
n = 165	*BCa 95%CI*	*Lower*	27.82	5.29	4.22
	of Mean	*Upper*	41.41	5.59	4.46
	SD		47.76	0.93	0.78
	Min		0.39	2.75	2.58
	Max		338.98	7.00	6.08
Multiword	*Mean*		8.74	4.42	4.43
Units	*BCa 95%CI*	*Lower*	1.89	4.01	4.15
n = 24	*of Mean*	*Upper*	18.37	4.83	4.70
	SD		24.19	0.96	0.68
	Min		0.02	3.00	2.92
	Max		114.00	6.21	5.58

Note. *Frequency* is based on occurrence per million words. Usefulness and difficulty were rated on a scale from 1 ("*Not worth"; "Basic/Easy"*) to 7 ("*Indispensable"; "Advanced/Difficult"*).

Table 2

Bootstrapped Pearson's Correlations of 189 Target Words

Criteria		*r*	*p*	R^2	*BCa 95% CI*	
					Lower	*Upper*
All Words (*n* = 189)						
Frequency	Usefulness	.470***	< .001	.221	.395	.567
Frequency	Difficulty	-.395***	< .001	.156	-.482	-.304
Usefulness	Difficulty	-.322***	< .001	.104	-.428	-.211
Single Words (*n* = 165)						
Frequency	Usefulness	.441***	< .001	.194	.358	.539
Frequency	Difficulty	-.420***	< .001	.176	-.508	-.327
Usefulness	Difficulty	-.340***	< .001	.116	-.451	-.218
Multiword Units (*n* = 24)						
Frequency	Usefulness	.556**	.005	.309	.216	.854
Frequency	Difficulty	-.058	.787	.003	-.264	.086
Usefulness	Difficulty	-.254	.232	.065	-.560	.065

Note. $^{*}p < .05$.

$^{**}p < .01$.

$^{***}p < .001$.

Table 3

Bootstrapped Descriptive Statistics of Three Clustering Solutions for 189 Target Words

Cluster		Frequency				Usefulness				Difficulty			
				BCa 95% CI of Mean				*BCa 95% CI of Mean*				*BCa 95% CI of Mean*	
No.	*N*	*Mean*	*SD*	*Lower*	*Upper*	*Mean*	*SD*	*Lower*	*Upper*	*Mean*	*SD*	*Lower*	*Upper*
Six-Cluster Solution													
1	35	9.53	9.77	6.46	12.70	4.67	0.46	4.52	4.82	3.97	0.52	3.79	4.14
2	23	3.37	6.58	1.37	6.11	3.48	0.39	3.32	3.64	4.89	0.42	4.71	5.07
3	60	19.94	16.71	15.87	24.21	5.39	0.59	5.24	5.54	5.09	0.37	5.00	5.18
4	55	32.62	19.04	27.60	37.53	6.10	0.42	5.99	6.21	3.82	0.48	3.69	3.94
5	13	124.71	32.69	109.19	142.42	6.25	0.49	5.98	6.53	3.59	0.61	3.25	3.92
6	3	279.96	58.65	240.87	319.06	6.54	0.19	6.42	6.67	3.14	0.24	3.07	3.32
Five-Cluster Solution													
1	35	9.53	9.77	6.47	12.78	4.67	0.46	4.52	4.82	3.97	0.52	3.78	4.14
2	23	3.37	6.58	1.37	6.04	3.48	0.39	3.31	3.65	4.89	0.42	4.72	5.06
3	60	19.94	16.71	15.88	24.23	5.39	0.59	5.24	5.54	5.09	0.37	5.00	5.18
4	55	32.62	19.04	27.65	37.57	6.10	0.42	5.99	6.21	3.82	0.48	3.68	3.94
5	16	153.82	72.32	123.39	188.67	6.31	0.46	6.07	6.53	3.59	0.61	3.23	3.78
Four-Cluster Solution													
1	58	7.09	9.11	4.88	9.48	4.20	0.73	4.01	4.39	4.34	0.66	4.16	4.51
2	60	19.94	16.71	15.92	24.19	5.39	0.59	5.24	5.54	5.09	0.37	5.00	5.18
3	55	32.62	19.04	27.57	37.64	6.10	0.42	5.99	6.21	3.82	0.48	3.68	3.94
4	16	153.82	72.32	122.18	190.68	6.31	0.46	6.07	6.55	3.50	0.58	3.22	3.78

Note. The changes in the number of words in different cluster groups for different clustering solutions revealed the process of grouping words from six to five and four clusters. *Clusters 5* and *6* in the *Six-Cluster Solution* were combined into *Cluster 5* in the *Five-Cluster Solution*, and based on this, *Clusters 1* and *2* in the *Six-* and *Five-Cluster Solutions* were further combined into *Cluster 1* in the *Four-Cluster Solution*.

Table 4

Clusters in the Five-Cluster Solution

Cluster	Group	*n*	Frequency	Usefulness	Difficulty	Example
1	$Freq_{Lo2}/Use_{Lo2}/Diff_{Mid}$	35	1.92	4.00	4.33	*skyrocket*
2	$Freq_{Lo1}/Use_{Lo1}/Diff_{Hi2}$	23	0.98	3.15	4.83	*incinerate*
3	$Freq_{Mid}/Use_{Mid}/Diff_{Hi1}$	60	21.23	4.67	5.46	*erosion*
4	$Freq_{Hi2}/Use_{Hi2}/Diff_{Lo2}$	55	32.62	6.10	3.82	*ethnicity*
5	$Freq_{Hi1}/Use_{Hi1}/Diff_{Lo1}$	16	153.82	6.31	3.59	*consequence*

Note. *Frequency* is based on occurrence per million words. Usefulness and difficulty were rated on a scale from 1 ("*Not worth"; "Basic/Easy"*) to 7 ("*Indispensable"; "Advanced/Difficult"*).

Footnotes

[1] The *Six-Cluster Solution* is identical to the *Five-Cluster Solution* except that *Cluster 5* in the *Five-Cluster Solution* was divided into a set of 13 words and another set of three words in the *Six-Cluster Solution* (see Table 3). Therefore, we could gain largely the same information by focusing the analyses on the *Four-* and *Five-Cluster Solutions*.