Vocabulary Learning and Instruction
Volume 8, Issue 1, March 2019
http://vli-journal.org

Evaluating Corpora with Word Lists and Word Difficulty

Brent A. Culligan
Aoyama Gakuin Women's Junior College
https://doi.org/10.7820/vli.v08.1.Culligan

Abstract

This study examines the application of an IRT analysis of words on lists including the General Service List (GSL), New General Service List (NGSL), Academic Word List (AWL), New Academic Word List (NAWL), and TOEIC Service List (TSL). By comparing line graphs, density distribution graphs, and boxplots for the average difficulty of each word list to related lists, we can get a visualization of the data's distribution. Japanese EFL students responded to one or more of 84 Yes/No test forms compiled from 5,880 unique real words and 2,520 nonwords. The real words were analyzed using Winsteps (Linacre, 2005) resulting in IRT estimates for each word. By summing the difficulties of each word, we can calculate the average difficulty of each word list which can then be used to rank the lists. In effect, the process supports the concurrent validity of the lists. The analysis indicates the word family approach results in more difficult word lists. The mean difficulties of the GSL and the BNC_COCA appear to be more divergent and more difficult particularly over the first 4000 words, possibly due to the use of Bauer and Nation's (1993) Affix Level 6 definition for their compilation. Finally, just as we should expect word lists for beginners to have higher frequency words than subsequent lists, we should also expect them to be easier with more words known to learners. This can be seen with the gradual but marked difference between the different word lists of the NGSL and its supplemental SPs.

Key words: IRT, word difficulty, corpus validity, measurement, vocabulary testing, Yes/No test

1. Measuring Word Difficulty Using IRT

At present, while there is a healthy debate on what constitutes a word for counting purposes, there are few ways to describe and evaluate the products of corpus analysis, whether their consistency or difficulty. In this study, I will look at one method to compare word lists derived from frequency analysis of corpora using word difficulty. Under this approach, word difficulty is measured using item response theory (IRT) applied to responses on Yes/No tests. IRT is a statistical model that estimates the probability of a person getting an item correct. More formally, IRT attempts to describe and predict the relationship between a random person and a random item, specifically how the person will respond to the item, given the ability of the person and the difficulty of the item. IRT posits that the

Please cite this article as: B. A. Culligan. (2019). Evaluating Corpora with Word Lists and Word Difficulty. *Vocabulary Learning and Instruction, 8*(1), 29–38. https://doi.org/10.7820/vli.v08.1.Culligan

higher the ability of the person, the higher the probability the person will respond correctly to the item. Conversely, the more difficult the item, the less likely the person is to correctly respond. For example, we would not expect a beginning English as a Foreign Language (EFL) student to know the word *residue*, but we would be very surprised if an advanced learner did not know the word *room*. If we know both the difficulty of the item and the ability of the person, we can predict the probability of the person correctly responding to a given item. IRT values are measured in logits, which can range from $-\infty$ to $+\infty$ but are more generally constrained to −7 to +7 for vocabulary difficulty. For word difficulties, the higher the logit, the more difficult a word is. For example, *room* at −6.43 is far easier than *elicit* at 5.38.

By summing the difficulties of each word, we are able to find the average difficulty of the word list. In this way, we can assess and rank word lists by difficulty. If we know the ability of a student, we will be able to ascertain how many of the words will be known to the student much more accurately than by testing a stratified random sample of the words in the list. For example, Nation's Vocabulary Size Test samples one word from 100, so to estimate the number of words known for each thousand-word band, they test 10 items. However, by using IRT estimates for word difficulty, we can get a better estimate of how many words are known by summing the probabilities of each word being known on the word list.

The purpose of this study is the application of the findings from the IRT analysis of the large data set to the assessment of the difficulty of a word list. The research question is to investigate what the average IRT difficulty estimates for the words on the General Service List (GSL), New General Service List (NGSL), Academic Word List (AWL), New Academic Word List (NAWL), and TOEIC Service List (TSL) can tell us about the use and validity of these lists?

2. Method

2.1. Participants

In this study, over 1,200 students from various schools responded to one or more of 84 different Yes/No test forms. Participants were primarily first and second year students enrolled in either a 2-year women's junior college or 4-year coed universities in Japan.

2.2. Materials

2.2.1. Yes/No tests

Real words. The real words from this study were compiled from word lists that are commonly found in the ESL/EFL literature. First to be considered were those lists that can be seen as useful for general English education. These lists were derived from corpora that tried to incorporate a diversity of genres and modalities. The first important list was West's (1953) *A General Service List of English Words*. This list of approximately 2,000 words has seen many revisions. As West claimed that "no attempt has been made to be rigidly consistent in the method

used for displaying the words: each word has been treated as a separate problem, and the sole aim has been clearness" (p. viii), his definition of what to count as a word was not explicitly stated. However, from the examples given within the text, he includes the frequency counts of the word used in different parts of speech and its inflections. Thus, we have nouns, verbs, and verbals (which include participles used as nouns and adjectives) contributing to the final frequency counts. For example, in West's entry for *feel* (p. 178–179), the total count for *feel* includes the verb (it feels soft) and the noun (the feel of silk), as well as the gerund *feeling* and its plural *feelings* (you've hurt my feelings). The version of the GSL used here was Bauman and Culligan's (1995) adaptation using Bauer and Nation's (1993) Word Family Level 3 affix definitions and ranked according to frequency counts from the Brown Corpus (Frances & Kučera, 1982). This resulted in a list of 2,284 words. Next examined was one of the updates to the GSL, the 2,801-word NGSL (Browne, Culligan, & Phillips, 2013a). The NGSL uses an extended lemma based on West's GSL for its definition of what is a word. Three other lists were derived from the Brown Corpus, the British National Corpus (BNC), and the BNC_COCA. The BNC_COCA (Nation, 2012) was designed by Paul Nation and colleagues and is an amalgamation of his BNC list (Nation, 2006) and data from the Corpus of Contemporary American English (COCA). It comes with Nation's Range software and is compiled into 1000-word bands using Bauer and Nation's Affix Level 6 definition (1993). For the BNC (Leech, Rayson, & Wilson, 2001), all words with a frequency of ten or more were used.

Second to be considered are the special purpose lists. The first of these is the long-established AWL (Coxhead, 2000) and the newer 963-word NAWL (Browne, Culligan, & Phillips, 2013b), followed by an examination of two newer lists, the 1,754 words Business Service List (BSL) and the 1,259 words TSL (Browne & Culligan, 2016a, 2016b). The AWL uses Bauer and Nation's (1993) Affix Level 6 definitions, while the NAWL, BSL, and TSL use the extended lemma. The list was then sorted, and all duplicates were eliminated, leaving 5,880 unique words.

Nonwords. The source of the 2,520 nonwords was the ARC database of nonwords (Rastle, Harrington, & Coltheart, 2002a, 2002b), which provided a list of nonwords with information on many characteristics including the number of neighbors. Nonwords that appeared to be inflected, as well as words that had too many neighbors with real words, were eliminated, thus providing an ideal source for nonwords.

2.2.2. Test forms

Each of the 84 test forms created for this study consisted of 70 real words and 30 nonwords. The 5,880 words and 2,520 nonwords were arranged in order of number of letters and randomly assigned to one of 120 tests. The words and nonwords on each form were then sorted by random number to determine their order on the test form. Each test form was assigned a unique test identification number (TestID) that was converted to a scanner-readable bar code. On the test, the 100 randomly ordered words and nonwords were arranged in four columns, with each word item directly followed by a bubble font version of a Y or an N.

2.3. Procedure

The procedure for administering the Yes/No tests was simple and efficient. The tests were given at the beginning of each class. The whole procedure for each test took less than 10 minutes. The number of test forms taken by each student ranged from one to five. The tests were staggered to ensure common items among the population.

The tests were scanned using the Remark Office OMR software (*Remark Office OMR*, 2000), resulting in a data file composed of a line for each test. Each line consisted of 102 comma-separated variables consisting of the student number, a TestID, and responses to 100 items, with Y for a *Yes* response, N for a *No* response, or an X if the item was omitted.

3. Analysis

To obtain estimates for difficulty based on IRT, the data were analyzed using Winsteps (Linacre, 2005). Before analysis, individual test forms were eliminated based on false alarms rates, as Meara (2010) suggests that proportions of 0.50 or above render a test unreliable. After individual test forms were eliminated, the resulting test data set was converted to where each line of data represented the students' responses to all items on all test forms, and then analyzed by Winsteps. This resulted in IRT estimates for each of the tested words.

4. Results

Initial analysis of the data resulted in 5880-word difficulty estimates. A total of 554 words received perfect scores where all students who saw the word said they knew it, while 75 received zero scores where no students indicated knowledge. Winsteps uses a different algorithm that depends on the responses of the students who take similar items to compute the perfect and zero scores (Linacre, 2005).

The first research question looked for the average difficulty of various word lists. All data were reported in logits, the log of the odds unit. The lists used for general ESL and EFL programs were first compared. The GSL (West, 1953) had an overall mean of −1.86 logits (SD = 2.20), while the NGSL was slightly easier (M = −1.92, SD = 1.97) (see Figure 1). The violin plots in the figure show boxplots surrounded by the density distributions of the difficulties. The boxplots show the median, the 25 and 75 percent quantiles, and the minimum and maximum. The black dots represent extreme scores. The peak of the density distributions represents the mode, the most common score in the distribution. We can see that the distributions are relatively similar in the bottom half but differ toward the top. However, the graphic shows that the median score, the 25th percentile, and the 75th percentile of the GSL are very similar to the NGSL.

The first 1,000 words of the GSL had a mean word difficulty of −3.22 (SD = 1.56), while the second 1,000-word list had an average of −0.79 (SD = 2.03), a difference of 2.43 logits. The mean of the last 284 words was −0.83 (SD = 2.10), which is slightly easier than the previous thousand. For the same breakdown, the NGSL had means of −3.30, −1.59, and −0.62 (SD = 1.54, 1.68, and 1.71), respectively.

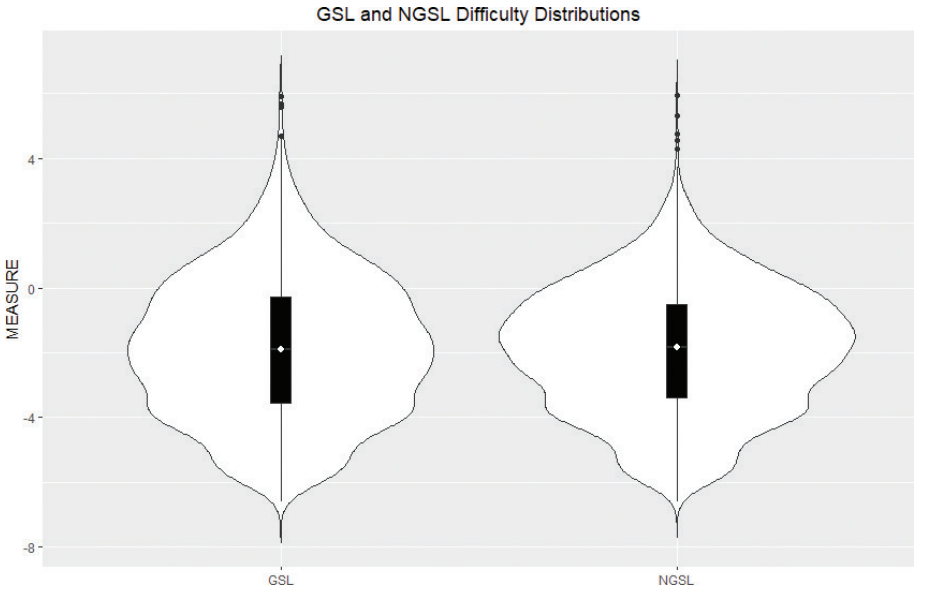

Figure 1. The Violin Plots (Boxplot and Difficulty Distribution) for the GSL and NGSL.

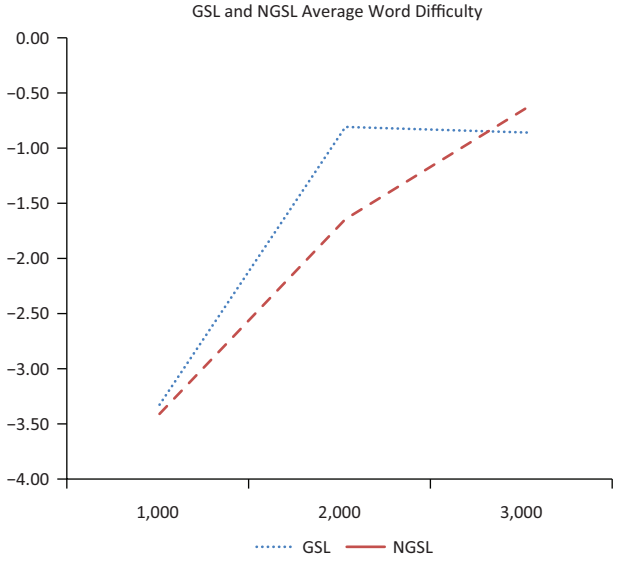

Figure 2. GSL and NGSL Word Difficulties by Band.

Graphically this is represented in Figure 2, where we can see the slight dip in the line graph for the GSL. Under the assumption that less frequent words tend to be more difficult, we would expect to see this line continue to incline upward given that these words are less frequent according to the frequency data from the BNC.

Next, we will look at the distribution of other lists derived from general corpora, specifically the Brown Corpus, the BNC, and the BNC_COCA. As can be seen in Figure 3, the BNC, Brown, and NGSL track fairly similarly, particularly over the first 4,000 words. However, the GSL and the BNC_COCA appear to be more divergent and more difficult. One possible explanation for the difference in overall difficulty is that the BNC_COCA uses Bauer and Nation's (1993) Affix Level 6 definitions. By including derivations with high frequencies with their base form, they are effectively removed from the lists. This means that in each band, and compounded subsequently, more words with lower frequencies will be included. For example, *government* is not found on the BNC_COCA as it is part of the *govern* word family, but in the BNC, it is in the first thousand, while *govern* is found in the third thousand. This concentration of words has an effect on base words as well. The third thousand of the BNC and Brown include words such as *rabbit* and *slight*, which are found in the BNC_COCA first thousand. The BNC_COCA third thousand includes words like *discriminate* and *affirm* that appear on the BNC's sixth and seventh thousand lists. For these reasons, we can see that how we define the lexical unit of counting has many ramifications for word lists extracted from corpus analysis.

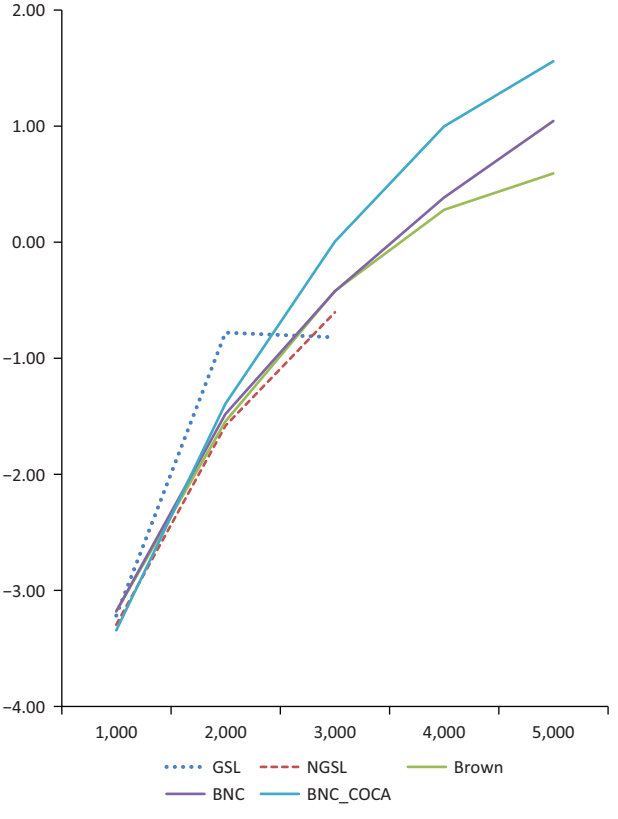

Figure 3. Growth in Average Word Difficulty.

With the introduction of the University Word List (Xue & Nation, 1984) and its replacement with the AWL (Coxhead, 2000), we see the beginnings of the use of Special Purpose (SP) vocabulary lists to supplement the GSL. Many such lists proliferated, such as a Medical Academic Word List (Wang, Liang, & Ge, 2008), a nursing list (Yang, 2015), and an engineering list (Ward, 2009). The NAWL was designed to supplement the NGSL in the same way the AWL supplemented the GSL. Figure 4 shows the distribution of the original AWL ($M = -0.12$, $SD = 2.07$) and more recent iteration, the NAWL ($M = 0.85$, $SD = 2.05$). The mean score of the NAWL is almost one logit more difficult than the AWL. From the plots, we can also see from the length of the tails that there are more extreme scores in the NAWL, particularly for words over six logits in difficulty. We can also see that the median of the AWL is clearly below the median and mode of the NAWL, thus affirming that the majority of the words in the NAWL lie above the mean, median, and mode of the AWL.

To get a more complete picture, it is necessary to see how the special purpose list fits with the base word lists. In Figure 5, we can see density distributions of the word difficulties for the GSL and the AWL. For this graph, the second word list of the GSL contained 2,284 words. As can be seen from the violin plots, while there seems to be a good separation between the first and second thousand bands of the GSL, there is a considerable overlap with word difficulty between the GSL 2 ($M = -0.80$, $SD = 2.19$) and the AWL ($M = -0.12$, $SD = 2.07$), with the box plots showing similar medians and the density distributions displaying similar modes. Visually, they appear to be occupying similar space on the difficulty spectrum.

In Figure 6, the NGSL and its supplemental SP word lists show a gradual, but much more marked difference between the different lists when compared to

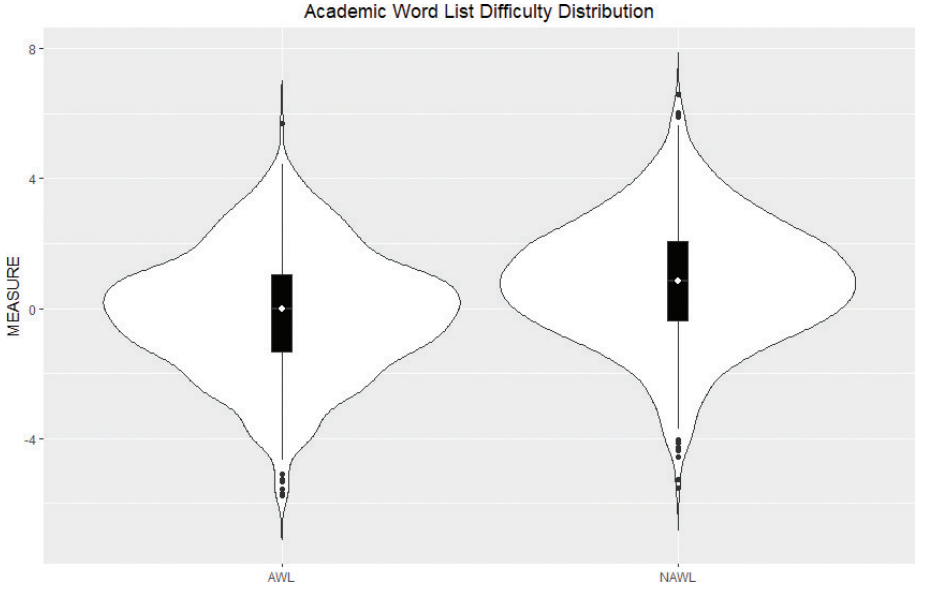

Figure 4. Violin Plots of Word Difficulty for Two Special Purpose Academic Word Lists.

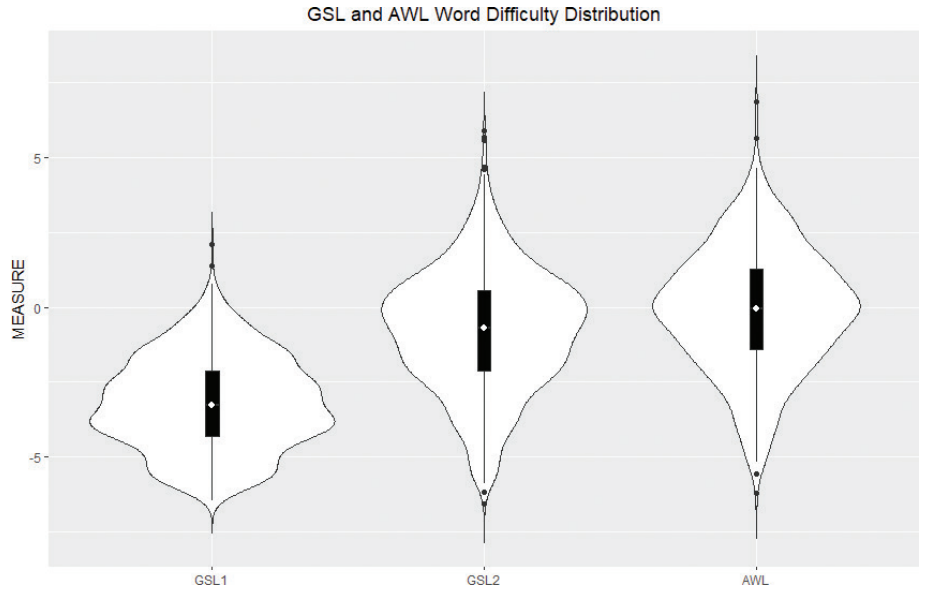

Figure 5. Violin Plots of the First and Second Thousand Words of the General Service List and the Academic Word List.

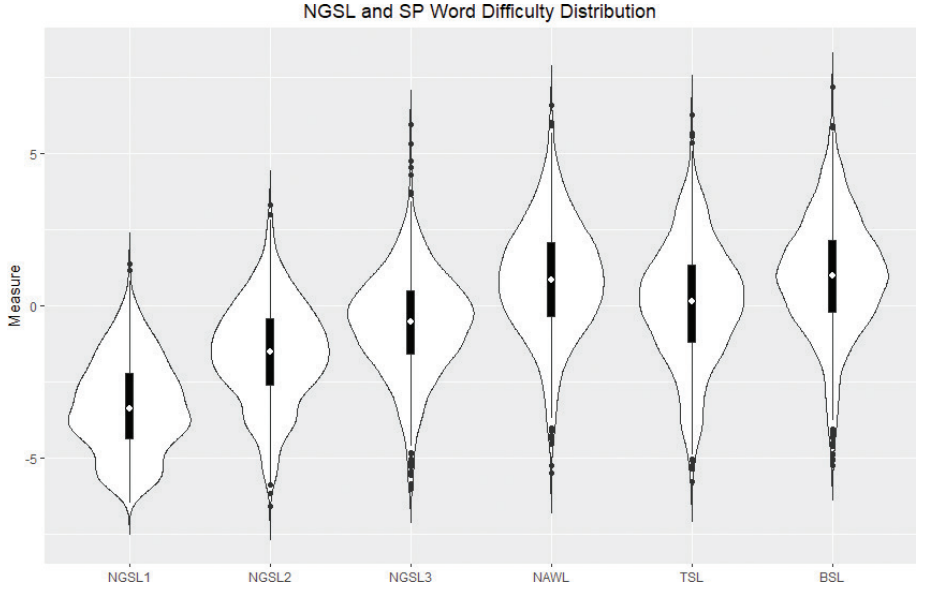

Figure 6. Density Distributions and Boxplots of the New General Service List and the New Academic Word List.

Figure 5. We can see that both the medians and the mode for each distribution are below their neighbor in the majority of cases. There are fewer differences between the SPs, but this is to be expected as the SPs are not mutually exclusive.

However, we can see some differences between the related SPs, the TSL and the BSL, which both deal with business type discourse. As the TOEIC is a test of business communication, this corpus is more limited in scope than the Business Corpus that covers a wide range of situations, genres, and text types. This is reflected in the increased difficulty of the BSL compared to the TSL.

5. Conclusion

This study looked at the application of difficulty to word lists produced from corpora. By comparing line graphs, density distribution graphs, and boxplots for each word list with related lists, we can get a better picture of how the data are distributed. This allows us to assess the validity of our word lists as well as investigate how well they work in conjunction with each other. Just as we should expect word lists for beginning courses to have higher frequency words than subsequent lists, we should also expect them to be easier, with more words known to the learners. In effect, the process described in this article provides a way to support the concurrent validity of our word lists. Finally, this analysis seems to indicate that using a word family approach, particularly using word affixation beyond Level 2, will result in more difficult word lists as was evidenced by the increased difficulty at successive thousand-word bands of the BNC_COCA based on Bauer and Nation's (1993) Affix Level 6 (see Figure 3). Similarly, the GSL based on Affix Level 4 showed the second thousand to be more difficult than those of the word lists using West's extended lemma approach.

References

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*, 253–279. doi:10.1093/ijl/6.4.253

Bauman, J., & Culligan, B. (1995). *About the general service list*. Retrieved from http://jbauman.com/aboutgsl.html

Browne, C., & Culligan, B. (2016a). *Business service list 1.01*. Retrieved from http://www.newgeneralservicelist.org/

Browne, C., & Culligan, B. (2016b). *TOEIC service list 1.1*. Retrieved from http://www.newgeneralservicelist.org/

Browne, C., Culligan, B., & Phillips, J. (2013a). *A new general service list (1.01)*. Retrieved August 12, 2016, from http://www.newgeneralservicelist.org/

Browne, C., Culligan, B., & Phillips, J. (2013b). *New academic word list 1.0*. Retrieved from http://www.newgeneralservicelist.org/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. doi:10.2307/3587951

Frances, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow: Longman.

Linacre, J. M. (2005). *WINSTEPS Rasch measurement computer program (Version 3.59)*. Beaverton, OR: Winsteps.com.

Meara, P. (2010). *EFL vocabulary tests (2nd ed.)*. Swansea: University of Wales, Centre for Applied Language Studies. Retrieved from http://www.lognostics.co.uk/vlibrary/meara1992z.pdf

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *36*(1), 59–82. doi:10.3138/cmlr.63.1.59

Nation, P. (2012). The BNC/COCA word family lists. Retrieved August 24, 2018, from https://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

Rastle, K., Harrington, J., & Coltheart, M. (2002a). 358, 534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology*, *55A*(4), 1339–1362. doi:10.1080/02724980244000099

Rastle, K., Harrington, J., & Coltheart, M. (2002b). 358, 534 nonwords: The ARC nonword database. Retrieved June 4, 2003, from www.maccs.mq.edu.au/~nwdb/nwdb.html

Remark Office OMR. (2000). Version 5. Principia.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442–458. doi:10.1016/j.esp.2008.05.003

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, *28*(3), 170–182. doi:10.1016/j.esp.2009.04.001

West, M. (1953). *A general service list of English words*. London: Longmans, Green and Co.

Xue, G., & Nation, P. (1984). A university word list. *Language Learning and Communication, 3*, 215–229. Retrieved from https://www.victoria.ac.nz/lals/about/staff/Publications/paul-nation/1984-Xue-UWL.pdf

Yang, M.-N. (2015). A nursing academic word list. *English for Specific Purposes, 37*, 27–38. doi:10.1016/j.esp.2014.05.003