Estimating the Likelihood of Words Being Known with Corpus Analysis and K-Means Clustering Algorithm

Tong Zhu[1]
Tong.Zhu@nottingham.edu.cn

Derek Irwin[1]
Derek.Irwin@nottingham.edu.cn

Yanhui Zhang[1]
Yanhui.Zhang@nottingham.edu.cn

Renjie Wu[2]
Renjie.Wu@nottingham.edu.cn

Xiaoyi Jiang[3]
Xiaoyi.Jiang@nottingham.edu.cn

[1]School of Education and English, Faculty of Humanities and Social Sciences, University of Nottingham Ningbo China
[2]School of Computer Science, Faculty of Science and Engineering, University of Nottingham Ningbo China
[3]Department of Mathematical Sciences, Faculty of Science and Engineering, University of Nottingham Ningbo China

Abstract

This study investigates the extent to which corpora-based frequency indices predict the likelihood of words being known. Five hundred and twenty ESL students in China's top-tier universities participated in this study. Their knowledge of target words to a meaning-recall level represented their likelihood of knowing these words. The target words were 932 content words retrieved from a text, of which the lexical coverage and readability level were appropriate for the participants. An online test was developed for the target words and administered to the participants. MATLAB was used to perform k-means clustering for participants' answers and classify the likelihood of words being known into the most appropriate number of clusters. SPSS was used to perform the Kruskal-Wallis test Spearman correlation, and ordinal logistic regression. In addition to the classification of the likelihood of words being known, results showed significant differences and moderate correlations between corpora-based frequency indices and the classification. Moreover, base words' frequency ranks on Nation's (2012) BNC/COCA list were found to best correlate with and predict the likelihood of words being known. Future research is recommended to extend this study by classifying more words' likelihood of being known to more learners at various levels of ESL proficiency.

1 Introduction

The likelihood of words being known is usually indicated by corpus analysis. Corpus linguistics has found that most words (e.g., *tyke*, *tantrum*) are low-frequency words and only occur limited times in corpora; on the contrary, around 2000 words (e.g., *have*, *think*) are high-frequency words and constitute 70% to 90% of texts (Dang et al., 2022; Dang and Webb, 2020; Nation, 2006). For the past decades, word frequency (i.e., how often a word appears in discourse) has been a key predictor of the likelihood of words being encountered, and thus being processed and known by learners (Ellis, 2002; Edwards and Collins, 2011; Horst and Collins, 2006; Schmitt et al., 2001, 2021). Studies measuring the vocabulary knowledge of L2 learners in different contexts (e.g., Henriksen and Danelund, 2015; Laufer, 1998; Matthews and Cheng, 2015; Nguyen and Webb, 2017; Stæhr, 2008; Webb and Chang, 2012) also showed that learners knew more high-frequency words than those at lower-frequency levels. Therefore, corpora-based frequency indices have generally been taken as a reasonable proxy to indicate the likelihood of words being known, and innovative methods have been developed to profile word frequency in corpora (Dang et al., 2022; Huang et al., 2022).

However, there may be significant differences between corpora-based frequency indices and the likelihood of words being known. It is true that learners typically know more words in high-frequency bands than lower-frequency bands, yet some words from lower-frequency bands were found to be well-known by L2 learners. For example, although *pencil* and *blackboard* are not necessarily frequent words in general English, they are important for the classroom context and are likely to be known by English as a Second Language (ESL) learners. Also, politeness terms may be infrequent in corpora, but are likely to be known, and profanity terms are generally absent from teaching materials, though frequent in some registers, and because of their taboo nature may be likely to be known among some students. Given that learners encounter specific words in their specific context, both word frequency and the likelihood of words being known should vary learner by learner (Laufer and Nation, 2012). Therefore, word frequency should be seen more as a useful indication rather than a prescription.

Recent research supports that corpora-based frequency indices imperfectly predict what learners might have learned (Brysbaert et al., 2021; Dang et al., 2022; Schmitt et al., 2021); other factors such as cognates (Schmitt et al., 2021), register (Brysbaert et al., 2021), learner perception of word usefulness and difficulty (He and Godfroid, 2019), and lexical variables (i.e., "orthographic, phonological, morphological, syntactic, and semantic characteristics" of words (Hashimoto and Egbert, 2019, p. 840)) were also found to play an influential role in the likelihood of words being known. These factors suggest a need for estimating the likelihood of words being known through direct tests of learner knowledge.

Although large-scale vocabulary tests have been administered to assess learner knowledge and explore its connection with corpora-based frequency indices, to the best of our knowledge, previous studies have neither classified words into specific categories indicating the likelihood of being known, nor have they explored the connection between different lexical units' (i.e., base word, lemma, and word type) corpora-based frequency indices and their likelihood of being known. In the present study, we hope to fill this gap by making the first attempt at applying the k-means clustering algorithm to classify words into the most appropriate number of categories, each of which indicates a specific degree of likelihood of being known. Furthermore, corpus results were combined with empirical evidence to verify the following hypotheses: a) there are significant differences between the likelihood of words being known and their different lexical units' frequency indices in corpora; and b) there are significant but small correlations between the likelihood of words being known and their different lexical units' frequency indices in corpora. To this aim, two research questions guided this study:

1) How can we classify the likelihood of words being known?
2) To what extent do corpora-based frequency indices predict the classification?

2 Methods

2.1 Participants

Five hundred and twenty students from top-tier universities participated in the present study. All of them were native speakers of Mandarin Chinese with English as their second language. They were 18-and-above years old and had studied English formally for at least 10 years in primary and secondary school. In the present study, they were informed of the study purpose and provided informed consent before filling out personal information forms and completing a test with their test time being recorded. Data from 44 students were discarded due to their extremely short test time. Therefore, 476 participants, ranging in grades from Year 1 undergraduate to Year 4 Ph.D. students, were included in the final data analysis of this study. According to their self-reported scores in standard English examinations, including National College Entrance Examination, College English Test Band-4 (CET-4) and/or Band-6 (CET-6), Test for English Major Band-4 (TEM-4) and/or Band-8 (TEM-8), International English Language Testing System (IELTS), and Test of English as a Foreign Language (TOEFL), their English proficiency was rated B1 to C1 level (intermediate to advanced) based on the Common European Framework of Reference (CEFR) for language proficiency. This study was approved by the ethics committee of our university.

2.2 Target Words

The target words were content words (i.e., nouns, verbs, adverbs, and adjectives) retrieved from the first chapter of *Harry Potter and the Philosopher's*

Stone (Rowling, 1997). The text was selected due to the following two reasons. First, the *Harry Potter* series has enjoyed wide popularity among China's university students, so the participants were likely to be interested in and familiar with the selected text. Second, students at top-tier universities in China who passed the Gaokao examination should reach an intermediate and above level of English. As the 95% lexical coverage of the text involves words from the most frequent 4000 words based on Nation's (2012) British National Corpus (BNC)/the Corpus of Contemporary American English (COCA) list, the text is readable and appropriate for the participants.

To retrieve content words from the text with their contextual Part of Speech (PoS), the Natural Language Toolkit (NLTK), an NLP algorithm library based on Python, was used. The algorithm function *word_tokenize* was used to retrieve word types while the algorithm function *lemmatize* and *snowball stemmer* from the NLTK class *WordNetLemmatizer* were respectively used to retrieve lemmas and base words from the text. The algorithm function *pos_tag* was used to match each word type's PoS with the context. However, NLTK algorithms sometimes produce erroneous results in terms of word recognition and PoS matching. For example, compound words are often recognized as two separate words, and the first noun of a noun-noun collocation is often labeled as an adjective. To correct these errors, all retrieved words were double-checked with words retrieved by lextutor, a widely used lexical profiler (available at https://www.lextutor.ca/vp/comp/). Since the lextutor retrieves words without their PoS information, we manually checked the in-text content words' PoS information and corrected all errors. In the end, all content words and their contextual PoS were inputted into an Excel sheet in the form of 932 word types, 819 lemmas, and 737 base words.

2.3 Test Instrument

As the smallest lexical unit, word type provides the most precise findings on vocabulary knowledge. The 932 word types with their PoS information were thus used for test items. The test had a total of 18 pages and was developed and administered online. An ethics form was presented on the first page. Participants who agreed with the form and clicked on the "I agree" button could take the test. The second page presented eight questions about participants' personal information, including their names, genders, schools, majors, grades, contacts, years of studying English, and standard English test results. The third page to the 18th page presented around 60 test items (i.e., retrieved word types and their contextual PoS) per page in multiple-choice format. Participants were asked to click on words they had never seen or were unsure about the meaning. In addition, one out of the 60 items per page was randomly selected to be presented in a combination of Yes/No and fill-in-the-blank formats, where participants were asked to write down the first language (L1) definition of the test item if they clicked on "*Yes, I know this word*".

2.4 Scoring

The 932 test items and 443,632 item responses were inputted into an Excel sheet. Each item had a score of one or zero for each participant. The 916 test items presented in the multiple-choice format were marked by participants' clicking behavior. If a participant clicked on a test item, the item would be marked as zero for the participant; otherwise, the item would be marked as one. The 16 test items presented in a combination of Yes/No and fill-in-the-blank formats were scored twice. First, the 16 items were scored based on participants' clicks. If a participant clicked on "*No, I don't know this word*" for an item, the item would be marked as zero for the participant; otherwise, the item would be marked as one. Next, the 16 items were scored based on participants' translations. If the definition of a test item was correctly given, the item would be marked by one; otherwise, the item would be marked by zero. To ensure the reliability of the scoring, two research assistants, whose IELTS scores respectively were 7.0 and 7.5 and reached the CEFR C1 level, examined and marked participants' translations by one or zero. SPSS software was used to calculate Cohen's Kappa statistics (k) for identifying the strength of agreement between the two markers. Results showed that their scorings reached a high Kappa value ($k = .98$, $p < .01$), indicating an almost perfect agreement between the two research assistants. After resolving the discrepancies, SPSS software was used again to calculate the test-retest reliability value for the 15,234 scores on the 16 test items and the internal consistency reliability value for the 443,632 scores on all test items. The Pearson correlation coefficient was .93 ($p < .01$), indicating a high test-retest reliability of the test. Cronbach's

alpha was .992 ($p < .01$), indicating a high internal consistency reliability of the test.

2.5 Data Analysis

MATLAB 2022a was used to perform k-means clustering, an unsupervised machine learning algorithm, for classifying test items into specific clusters of the likelihood of being known. As shown in Figure 1, the k-means clustering algorithm is an iterative grouping algorithm with the following steps (Pujianto et al., 2019):

Step 1: Initialize k-means parameters, including but not limited to determining the number of clusters and the initial value of k cluster centers.

Step 2: Assign each test item to the closest pre-defined k cluster center and calculate the Euclidean distance between each item and the center.

Step 3: Recalculate the centers with the current items of each cluster.

Step 4: Repeat Step 2 and Step 3 until there is no change between the previous centers and the new centers.

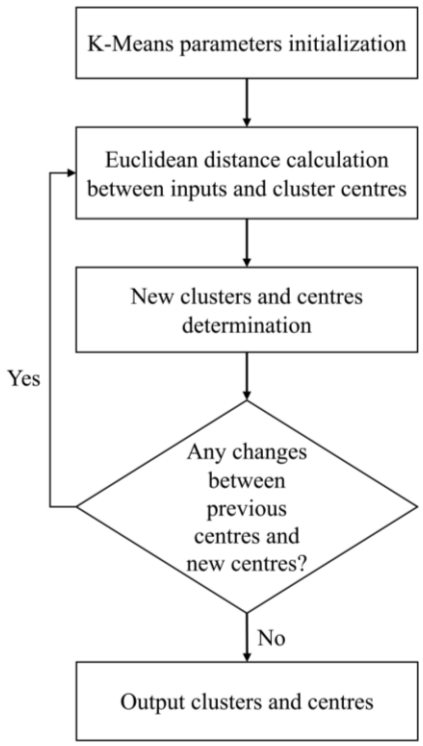

Figure 1: Flowchart of k-means clustering algorithm

Based on output centers and clusters, SPSS was used to perform the Kruskal-Wallis test and Spearman correlation for the clusters of the likelihood of test items (i.e., the 932 word types retrieved from the text) being known, their frequency counts in COCA, their lemmas' frequency counts and ranks in COCA, and their base words' frequency ranks on Nation's (2012) BNC/COCA list. Instead of the Pearson correlation performed previously, Spearman correlation was chosen this time due to the fact that clusters are ordinal variables rather than interval variables.

3 Results and Discussion

3.1 The Classification of The Likelihood of Words Being Known

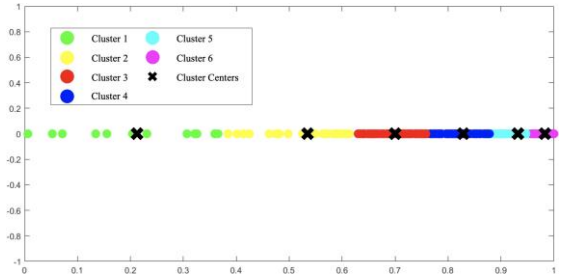

Figure 2: K-means clustering graph

After 10 iterations, the likelihood of test items being known was classified into six clusters (*epsilon* = .10, *average calculation error* = .05), and each test item was allocated to one of the six clusters. Figure 2 illustrates the k-means clustering results. Each color area represents one cluster with its center being highlighted by a black cross mark. As indicated by the x-axis, value 1 means that the test item is likely to be known by all participants, while value 0 means that the test item is likely to be known by no one. Accordingly, test items classified in cluster 1 are words least likely to be known, while test items classified in cluster 6 are words most likely to be known. Since the 6 clusters of the likelihood of test items being known were based on item responses of as many as 443,632, the results can counterbalance the potential influences of lexical variables and register but cannot be applied to a larger participant group. This is because learners with various levels of ESL proficiency may have different perceptions of word usefulness and word difficulty (He and Godgroid, 2019), which in turn affects the likelihood of words being known.

3.2 Corpora-Based Frequency Indices and The Likelihood of Words Being Known

We retrieved the frequency counts of the 932 word types in COCA (i.e., word type COCA count), the frequency counts and ranks of the 819 lemmas in COCA (i.e., lemma COCA count and rank), and the frequency ranks of the 737 base words in Nation's (2012) BNC/COCA list (i.e., base word k-value). The base word k-value and the clusters of the likelihood of words being known were two ordinal variables. Word type COCA count and lemma COCA count and rank were three scale variables, for which the Shapiro-Wilk test (S-W test) was performed since the sample size was small ($n < 5000$). As presented in Table 1, statistically significant S-W test results were found for the three scale variables, indicating that data did not meet normal distribution and the Kruskal-Wallis test is more applicable.

Statistically significant but small differences were found between corpora-based frequency indices and the clusters of the likelihood of words being known. As can be seen in Table 1, base word k-value ($f = .07$, $p < .01$), lemma COCA rank ($f = .10$, $p < .01$), lemma COCA count ($f = .01$, $p < .01$), and word type COCA count ($f = .02$, $p < .01$) were statistically significantly different from the clusters of the likelihood of words being known, but the differences were similarly small. Given that this study used word type as the lexical unit for test items, the similarly small difference of corpora-based frequency indices for base word, lemma, and word type supported the underlying assumption of different lexical units. Using base word as the lexical unit assumes that once learners know the meaning of a base word (e.g., *calculate*) and have some knowledge of morphology, they do not need to learn every single word in a language but instead can derive the meanings of word family members (e.g., *calculation*; *calculator*) from the base form (Vilkaite-Lozdiene & Schmitt, 2019). Similarly, the underlying assumption for lemmas is that once learners know a base word, they are likely to know its inflected forms and their PoS. However, as the smallest counting unit, word type assumes that learners have to learn words one by one. Therefore, although needing further justification with empirical evidence, we found that base word and lemma were more appropriate than word type in efficiently assessing word knowledge of learners who reached intermediate and above level of EFL proficiency.

Table 1 also showed statistically significant and moderate correlations between corpora-based frequency indices and the clusters of the likelihood of words being known. To be specific, base word k-value ($rho = -.69$, $p < .01$) and lemma COCA rank ($rho = -.64$, $p < .01$) were negatively correlated with the clusters, while lemma COCA count ($rho = .63$, $p < .01$) and word type COCA count ($rho = .54$, $p < .01$) were positively correlated with the clusters. As the likelihood of words being known increases with the number of clusters they belong to, the negative correlations could be explained by the fact that a higher k-value of a base word means that the base word belongs to a lower frequency level and thus is less likely to be known. Similarly, the higher rank of a lemma in COCA represents a lower likelihood of the lemma being known. In short, the larger the base word k-value or lemma COCA rank, the lower the likelihood of words being known. On the contrary, a larger lemma COCA count or word type COCA count represents the more frequent appearances of the words, and thus a higher likelihood of words being known.

To further examine the extent to which corpora-based frequency indices predict the clusters of the likelihood of words being known, the ordinal logistic regression was performed. As the likelihood ratio chi-square value was statistically significant, the regression model was effective. Nevertheless, the Odds Ratio (OR) was equal to one for the lemma COCA rank, lemma COCA count, and word type COCA count, indicating that the odds of being in a higher category for a one-unit increase in the three variables were the same as the odds of being in a lower category. In other words, a lemma's frequency rank in COCA, its frequency count in COCA, or a word type's frequency count in COCA had no effect on moving up or down the clusters of the likelihood of words being known. Base word k-value, on the other hand, was found to be a statistically significant predictor whose OR was .88. Therefore, base words' frequency ranks on Nation's (2012) BNC/COCA list had a negative effect on the likelihood of being in a higher cluster. Specifically, the odds of being in a higher cluster decrease by 11.83% with one-unit increase in the base word k-value.

	S-W Test	Cluster	N	Mean Rank	Kruskal-Wallis H	Cohen's *f*	Spearman Correlation
Base Word K-Value		1	12	785.79	452.01 ($p <$ .01)	.07	**-.69 ($p <$ .01)**
		2	34	816.50			
		3	74	723.24			
		4	77	708.56			
		5	154	561.41			
		6	581	349.49			
		Total	932				
Lemma COCA Rank	.63 ($p <$ .01)	1	12	815.83	376.64 ($p <$ .01)	.10	-.64 ($p <$ .01)
		2	34	811.66			
		3	74	744.03			
		4	77	690.29			
		5	154	599.84			
		6	581	338.74			
		Total	932				
Lemma COCA Count	.16 ($p <$ .01)	1	12	118.04	366.57 ($p <$ .01)	**.01**	.63 ($p <$ .01)
		2	34	122.44			
		3	74	194.07			
		4	77	247.82			
		5	154	334.51			
		6	581	592.50			
		Total	932				
Word Type COCA Count	.32 ($p <$ .01)	1	12	69.13	272.16 ($p <$ .01)	**.02**	.54 ($p <$ .01)
		2	34	220.62			
		3	74	235.06			
		4	77	264.32			
		5	154	356.36			
		6	581	574.59			
		Total	932				

Table 1: Kruskal-Wallis test and Spearman correlation result

Likelihood Ratio Chi-Square	Predictor	Coefficient	Standard Error	z	P	OR	95% Confidence Interval
404.586 ($p <$ .01)	Base Word K-Value	-.13	.03	-4.96	**< .01**	**.88**	.84-.93
	Lemma COCA Rank	.00	.00	-11.37	**< .01**	1.00	1.00-1.00
	Lemma COCA Count	.00	.00	.05	.96	1.00	1.00-1.00
	Word Type COCA Count	.00	.00	1.96	.05	1.00	1.00-1.00

Table 2: Ordinal logistic regression result

In summary, the present study found that corpora-based frequency indices were a crude proxy for estimating the likelihood of words being known. This finding is in line with Hashimoto and Egbert (2019), Schmitt et al. (2021), and Dang et al. (2022), who also reported small to moderate correlations between corpora-based frequency indices and learner-based word knowledge. Furthermore, among the four corpora-based frequency indices, base words' frequency ranks on Nation's (2012) BNC/COCA list were found to best correlate with and predict the classification of the likelihood of words being known. This may be explained by the fact that this list was made upon not only frequency indices but also subjective criteria. Several common spoken words (e.g., *goodbye*), weekdays, months, and numbers were perceived as simple words and were included in this list. It seems that corpora-based frequency indices should be used in conjunction with subjective criteria, such as perceptions of word usefulness and difficulty (He and Godfroid, 2019; Dang et al., 2022), for better estimating the likelihood of words being known.

4 Conclusion

Drawing upon frequency indices for word types, lemma, and base words from COCA and Nation's (2012) BNC/COCA list, as well as 443,632 item responses to 932 word types, this study supports the current trend in using learner-based word knowledge for estimating the likelihood of words being known (Brysbaert et al., 2020; Dang et al., 2022; Schmitt et al., 2021). As the first study to combine corpus analysis with empirical evidence while using k-means clustering algorithm, Kruskal-Wallis test Spearman correlation, and ordinal logistic regression, we extend EFL vocabulary understanding in several ways.

First, we challenge the theoretical convention of solely basing likelihood of words being known on corpora-based frequency indices, and support recent research that it should be estimated by learner-based word knowledge. Second, we suggest that EFL learners and teachers should combine corpora-based frequency indices with their subjective perceptions in vocabulary learning and teaching. Finally, we innovatively use the k-means clustering algorithm for classifying the likelihood of words being known into the most appropriate number of clusters. Such classification affords word selection for academic and pedagogical use, such as updating word lists, calibrating lexical pools, and personalizing lexical glosses. A recent meta-analysis concluded that personalized lexical gloss, although yet to be developed, has great potential in enhancing L2 vocabulary learning (Zhu et al., 2023). Classifying words' likelihood of being known can efficiently identify each individual learner's unknown words and prepare personalized lexical gloss for further exploration.

Nevertheless, this study is limited by the number of test items and the scope of participants. As a result, the clusters of the likelihood of being known should not be applied beyond the test items, or to EFL learners who have different education and language backgrounds from the participants of this study. Future research is recommended to a) update the k-means clustering algorithm with multiple appropriate resources (e.g., Nation's (2012) BNC/COCA list, Schmitt et al's (2021) KVL list) to non-linearly predict the clusters for all lemmas and word types in COCA; and b) replicate the present study with different words and diversified participants.

Acknowledgments

This study was supported by a Li Dak Sum Innovation Fellowship Grant and the Language and Pedagogy Lab at the University of Nottingham Ningbo China.

References

Brysbaert, M., Keuleers, E., and Mandera, P. 2021. Which words do English non-native speakers know? New supernational levels based on yes/no decision. *Second Language Research*, *37*(2):207–231. https://doi-org.ezproxy.nottingham.edu.cn/10.1177/0267658320934526.

Dang, T. N. Y., and Webb, S. 2020. Vocabulary and Good Language Teachers. In C. Griffiths and Z. Tajeddin (Eds.), *Lessons from Good Language Teachers* (1st ed., pp. 203–218). Cambridge University Press. https://doi.org/10.1017/9781108774390.019.

Dang, T. N. Y., Webb, S., and Coxhead, A. 2022. Relationships between lexical coverage, learner knowledge, and teacher perceptions of the usefulness of high-frequency words. *Foreign Language Annals*, *55*(4):1212–1230. https://doi.org/10.1111/flan.12663.

Edwards, R., and Collins, L. 2011. Lexical Frequency Profiles and Zipf's Law. *Language Learning*,

61(1):1–30. https://doi.org/10.1111/j.1467-9922.2010.00616.x.

Ellis, N. C. 2002. Frequency effects in language processing. Studies in Second Language Acquisition, 24(2):143–188. https://doi.org/10.1017/S0272263102002024.

Hashimoto, B. J., and Egbert, J. 2019. More Than Frequency? Exploring Predictors of Word Difficulty for Second Language Learners. *Language Learning*, *69*(4):839–872. https://doi.org/10.1111/lang.12353.

He, X., and 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.

Henriksen, B., and Danelund, L. 2015. Studies of Danish L2 learners' vocabulary knowledge and the lexical richness of their written production in English. In P. Pietilä, K. Doró, & R. Pipalová, *Lexical issues in L2 writing* (pp. 29–55). Cambridge Scholars Publishing. https://ebookcentral.proquest.com/lib/nottingham/detail.action?docID=4534718.

Horst, M., and Collins, L. 2006. From faible to strong: How does their vocabulary grow? *Canadian Modern Language Review*, *63*(1):83–106. https://doi.org/10.3138/cmlr.63.1.83.

Huang, L., Ouyang, J., and Jiang, J. 2022. The relationship of word processing with L2 reading comprehension and working memory: Insights from eye-tracking. *Learning and Individual Differences*, *95*:102143. https://doi.org/10.1016/j.lindif.2022.102143.

Laufer, B. 1998. The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics*, *19*(2):255–271. https://doi.org/10.1093/applin/19.2.255.

Laufer, B., and Nation, I. S. P. 2012. Vocabulary. In G. S. M & M. A, *The Routledge handbook of second language acquisition* (pp. 163–176). Routledge. https://ebookcentral.proquest.com/lib/nottingham/detail.action?docID=5655355.

Matthews, J., and Cheng, J. 2015. Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, *52*:1–13. https://doi.org/10.1016/j.system.2015.04.015.

Nation, I. S. P. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1):59–82. https://doi.org/10.3138/cmlr.63.1.59.

Nation, I. S. P. 2012. The BNC/COCA word family lists. Available at: https://www.lextutor.ca/cgi-bin/vp/comp/lists.pl?frame=bnc_coca_heads.

Nguyen, T. M. H., and Webb, S. 2017. Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3):298–320. https://doi.org/10.1177/1362168816639619.

Pujianto, U., Hidayat, M. F., and Rosyid, H. A. 2019. Text Difficulty Classification Based on Lexile Levels Using K-Means Clustering and Multinomial Naive Bayes. In *Proceedings of 2019 International Seminar on Application for Technology of Information and Communication (ISemantic)*, pages 163–170. https://doi.org/10.1109/ISEMANTIC.2019.8884317.

Rowling, J.K. 1997. *Harry Potter and the Philosopher's Stone.* London: Bloomsbury.

Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., and Kremmel, B. 2021. Introducing Knowledge-based Vocabulary Lists (KVL). *TESOL Journal*, *12*(4):e622. https://doi.org/10.1002/tesj.622.

Schmitt, N., Schmitt, D., and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1):55–88. https://doi.org/10.1177/026553220101800103.

Stæhr, L. S. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2):139–152. https://doi.org/10.1080/09571730802389975.

Vilkaitė-Lozdienė, L., and Schmitt, N. 2019. Frequency as a Guide for Vocabulary Usefulness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 81–96). Routledge. https://doi.org/10.4324/9780429291586-6.

Webb, S. A., and Chang, A. C.-S. 2012. Second Language Vocabulary Growth. *RELC Journal*, *43*(1):113–126. https://doi.org/10.1177/0033688212439367.

Zhu, T., Zhang, Y. H., and Irwin, D. 2023. Second and Foreign Language Vocabulary Learning through Digital Reading: A Meta-Analysis. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-11969-1.