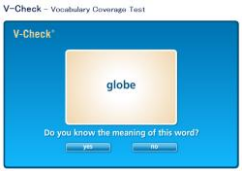
Word Difficulty, Word Frequency, and Spaced Repetition

Guy Cihi (2012)

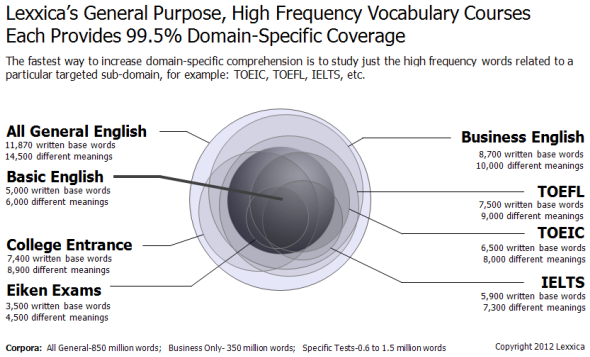
The frequency of a word's occurrence is often used to determine a word's rank order of importance to comprehension. It is widely agreed that the more frequently a word occurs, the more important it is to comprehension. The frequency of a word in text and speech, however, has not proven a reliable indicator of how widely known and well understood that particular word is. Lexxica's research lab has revealed that there is, in fact, low correlation ($corr(X,Y) = 0.6$) between general English word frequency and word difficulty. The words 'injured' and 'hurt' provide a good example: 'Injured' is a well known word to most junior high students across Asia, and yet 'hurt,' which is three times more frequent than 'injured,' is not well known to most college students in Asia. What's going on?

Lexxica R&D lab has developed a proprietary database of word difficulties for the 50,000 most frequent words in the English language. We have named our difficulty metric the "lexxit." For each of the 50,000 words in our database the lexxit is determined through statistical analysis of millions of "yes" "no" responses collected from hundreds of thousands of people who take our free V-Check lexical ability test. Lexxica first started offering the V-Check test free online in 2006. At that time we had just 6000 lexxits calibrated for the first 6000 most frequent words. Each V-Check test introduces five words for which we do not yet know the lexxit and in this way we have been able to collect sufficient response data to calibrate the lexxits for 27,000 words, and estimate the lexxits for another 23,000 making our combined total now 50,000 lexxits.

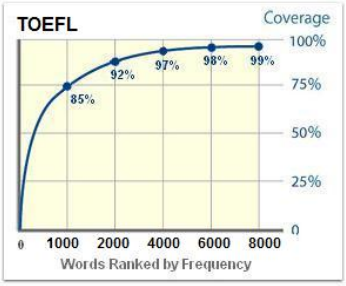


Just as words have different lexxits so, too, do people have different lexical abilities. The V-Check test employs a systematic patented process (CAT/IRT) to determine the lexical ability of each respondent. The free test measures the ability of each respondent against all of the 50,000 known lexxits. In our system, human lexical ability and lexxit word difficulty are measured along the exact same logarithmic scale. In our system a respondent determined to have lexical ability of 10.0 has a 50 percent probability of knowing any given word with a lexxit of 10.0. The same respondent will have a statistically higher probability of knowing a word with a lexxit of 5.0, and a statistically lower probability of knowing a word with a lexxit of 15.0.

Difficulty lexxits provide a practical and reliable way to identify which specific words are known to a learner, and which specific words are unknown. The findings make learning unknown vocabulary (and all factual knowledge) extremely efficient. Why waste time learning words you already know. Lexxica's V-Check lets you focus study time on only the most important words that you don't already know.



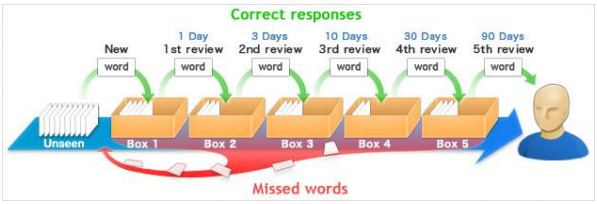
To determine which words (and facts) are important, Lexxica counts word frequencies in corpus as it best informs us which specific words will equip a learner for full comprehension. Most linguists now agree that 97 percent lexical coverage is the minimum target to achieve meaningful reading comprehension. Lexxica's Word Engine high speed learning system is designed to teach learners the vocabulary words required to attain 99 percent coverage of each domain.



Lexxica scans each corpus, for example our 1.25 million word TOEFL corpus, to mathematically determine the percentage of coverage that each different high frequency word provides for the corpus. In TOEFL, the word "office" provides 0.2 percent coverage of the corpus. Working with the rank-order list of all 16,736 different words that occur in our TOEFL corpus, we can add up all the individual item coverage percentages until the cumulative total reaches 99 percent. In the case of TOEFL it requires the 7,501 most frequent words to add up to 99 percent cumulative coverage.

Lexxica's V-Check test identifies each learner's lexical ability against all 50,000 words in our lexxit database. Once V-Check identifies the learner's ability, the system can easily determine which of the 7,501 TOEFL words that particular learner is likely to already know, and which specific TOEFL words he or she is unlikely to know. The Word Engine system can then create a personal target list of important, but unknown words for each learner. Lexxica's Word Engine always teaches the most frequently occurring unknown words first. Word Engine also employs a depth of knowledge assessment from V-Check to adjust the learner's starting point such that they begin by receiving about 65 percent partially known words, and 35 percent unknown words. As each learner progresses, he or she will move up into more and more unknown (higher lexxit) words. The Word Engine approach allows learners to develop confidence with the learning applications, and more importantly to review the correct meanings of the many high frequency words they only vaguely know.

The Word Engine uses spaced repetition to promote long-term retention for all new words. The 19th century German psychologist Dr. Hermann Ebbinghaus published the first research on human memory in 1895, and to this day all spaced repetition systems are based on his work. The first practical spaced repetition system based on Ebbinghaus' research was developed by Sebastian Leitner. It used cardboard boxes to sequence vocabulary flashcards over the increasing intervals of time described by Ebbinghaus. Lexxica co-founder Dr. Charles Browne, a respected advisor and teacher-trainer for the Japanese Ministry of Education, visited a high school in Kyoto where, remarkably, almost all of the students spoke fluent English. After reviewing their curriculum Dr. Browne concluded that the only significant difference at the school was their regular use of vocabulary flashcards and cardboard 'Lietner boxes.'



With spaced repetition, the learner must correctly identify each particular word at each different time interval, or the word gets sent back to the beginning. The main weakness of Leitner's original card boxes system was that it placed an enormous burden on teachers and students to prepare and manage thousands of paper flashcards across five cardboard boxes. Lexxica's online and mobile Word Engine digitizes and automates each student's flashcards and boxes, and provides corpus-based, personalized word lists for a variety of academic, general and business subjects.

If you have questions or comments please contact Guy Cihi at Lexxica R&D Ltd.
Email: gcihi (at) lexxica.co.jp

To experience the Word Engine please visit
wordengine.cn.com (China)
wordengine.jp (Japan)
wordengine.kr (Korea)

Lexxica R&D Ltd.
Tokyo, New York, Seattle