Measuring Vocabulary Size via Online Technology

Browne, Cihi and Culligan

Last updated March 2007

Summary

How many English words are there?

There are likely over two million English words in all forms when scientific terms are included, and likely over four million if organism and specie designations were to be included (Crystal, 1990).

Webster's Third New International Dictionary contains about 267,000 entries. Paul Nation classified 113,161 of those entries as word families (Nation, 1990). [We are happy to disclose here that Paul Nation advises Lexxica and the development of its services.]

The largest credible estimate we know of is from Henry Kucera. He has suggested the probable existence of some 375,000 English words, including proper words and special terms. He further suggested the 375,000 words would extend to about 600,000 English words in all forms based on his widely accepted ratio of 1 to 1.6 (Kucera, 1982).

How does Lexxica count words?

There are a variety of ways to count the words in the English language. Take for example the following six words:

Accept
Accepts
Accepting
Acceptable
Acceptance
Unacceptable

If we were to count these in terms of a "word family", there would be just one word, 'accept'. If we were to count in terms of lemmas, all 6 items would be counted. Which is correct? We believe the answer lies somewhere in between. Our preliminary findings indicate that the statistical item difficulty factors for 'accept', 'accepts' and 'accepting' are very close, whereas the statistical difficulties for 'acceptable', 'acceptance' and 'unacceptable', are all quite different. One hypothesis is that the brain treats these six items as four different Base Words: 'accept', 'acceptable', 'acceptance', and 'unacceptable'. As our database grows we will be able to identify with increasing accuracy how many discrete Base Words there are in any language.

What do we mean by Base Words?

By our reckoning, a Base Word is any word, or set of word forms, that the brain recognizes as one lexical unit. Base Words may manifest in multiple related word forms, as with 'accept', 'accepts', and 'accepting', or Base Words may manifest in just one form, as with the word 'the'. Base Words that manifest in multiple forms share a defining characteristic in our approach in that each form must have the same, or almost the same, statistical difficulty factor among a population. In practical terms the Base Word designation means that whenever a person indicates recognition of any one form of a Base Word, there can be a high degree of confidence that they will recognize all forms of that Base Word.

How many English Base Words do people know?

V-Check English Base Word recognition findings to date:

NOTICE: Lexxica systems are still early in their evolution. As of the date first indicated above, V-Check is certified for Japanese persons to assess English vocabulary knowledge up to the 6000th most important Base Word. This chart is provided solely to demonstrate how word recognition will differ between different population groups. It represents mean vocabulary sizes by population group however, any findings marked with an asterisk (*), are UNCERTIFIED PROVISIONAL findings generated from very limited samples.

Culture / Demographic	Average (mean) number of all known English Base Words
Japan / Age 21-25, C/U, M&F	**3,708**
Japan / Age 17-20, HS, M&F	2,984
Japan / Age 14-16, JHS, M&F	2,102
USA / Age 35+, M/PhD, M&F	42,732*
USA / Age 25-55, C/U, M&F	33,739*
USA / Age 17-20, HS, M&F	22,996*
USA / Age 14-16, ES, M&F	17,239*
China / Age 35+, M/PhD, M&F	12,854*
Korea / Age 21-25, C/U, M&F	4,244*
Taiwan / Age 21-25, C/U, M&F	4,294*

How many English Base Words do people need in order to communicate effectively?

Our methods and applications are lexical in nature - not grammatical or structural. Speaking from a lexical perspective, knowing an average of 19 out of every 20 words (95 percent coverage) of a written text is sufficient for effective comprehension. 95 percent coverage would permit a reader to comprehend the meaning without aid of a dictionary. The meanings of the 5 percent (or fewer) unrecognized words could be adequately grasped through context.

What do we mean by coverage?

The term coverage describes how many vocabulary words are known. Simply stated, “coverage” is a way of measuring and describing the amount of words in a text or spoken dialog that are known by the receiver. For reading, research indicates that knowing 19 of every 20 words, or 95 percent coverage, is the important threshold beyond which people can self-learn new words without the aid of a dictionary. Our research indicates that the 5,000 most important English Base Words are more than sufficient to “cover” 95 percent of general written English, and just 1500 most-important Base Words will effectively “cover” communication in spoken English.

Certain words tend to be better known among populations, and to occur more frequently in print than other words. Looking at frequency of occurrence, for example, ‘the’ is the most frequent word in the English language representing, or covering, about 7 percent of all the English words one is likely to ever encounter. Knowledge of the top 10 most frequent words represents, or covers, 25 percent of the words used in almost all written texts. Coverage then, generally describes the relationship between known vocabulary and the lexicon of a corpus. The chart below describes the relationship between high frequency English words and the well-known British National Corpus (BNC).

High Frequency Words	Percentage Coverage of BNC
1	7
10	25
100	50
1000	75
2000	85
3300	90
4000	95
6000	98
375,0000	100

These BNC findings are based on the word family method of counting. Lexxica organizes and counts words using a Base Word approach. Base Words are single citations that represent sets of related word forms. Base Words include standard inflected word forms and in some cases derived word forms. The more widely known word family method, described by Nation (1991), includes multiple derived word forms in each citation based on a fixed set of criteria and without regard for difficulty.

We have found that derived forms of words tend to vary widely in terms of difficulty. Lexxica hypothesizes that related word forms having the same measure of difficulty are being stored and processed similarly by the brain. Word forms that have different difficulties are likely being treated as different words. As a result, at the 95 percent level of coverage, the Base Word method will typically indicate about 25 percent more word citations than the word family method. For example BNC researchers have estimated that 4,000 words cover 95 percent of general texts. Lexxica estimates that 5,000 words are required. Admittedly there is tremendous overlap in the two approaches, and regardless of which is favored, it is highly recommended to make instruction of these most important words an integral part of any language program.

English is a remarkably efficient language with which people can easily survive and even thrive with limited vocabularies. Beyond the first several thousand important words, the remaining low importance words add tremendous, depth, flexibility and color to the language, but for most communications they are optional. However even low frequency words quickly become statistically important to comprehension and coverage when the subject matter concerned is a special purpose domain such as a professional vocation, an academic focus, or a career interest.

What is the significance of 95 percent coverage?

Research shows us that 95 percent coverage enables one to comprehend meaning without the aid of a dictionary, and coverage of less than 95 percent requires the use of a dictionary, and that coverage of less than 85 percent will generally defy comprehension regardless of dictionary use.

Following are sample paragraphs concerning two popular media figures. The first is shown at **67 percent coverage** where just 13.4 out of every 20 running words are recognizable. Selected words in the text have been scrambled to simulate the experience of reading unknown words. Try to figure out the meaning of this passage and/or identify the missing words:

> Brad Pitt told Marchilate mviswabe that he and Angelina Jolie will not be winplurtzd until the smorte to winplurtz is fromptes to bilps and plortes. Pitt, who trimpted the fitzleg of the smigteglar Bortslig fratmack, says, “Angie and I will consider gigrit the tonk when everyone else in the nonctron who wants to be winplurtzd is bleah.”

Lexxica co-founder Charles Browne's research has identified that 67 percent is the average coverage Japanese high school students have for their EFL textbooks. Regardless of the purpose or focus of any textbook, at 67 percent coverage, reading it will be nearly impossible.

Here below is the same paragraph shown at **95 percent coverage** where 19 out of every 20 running words are recognizable and only a few words are scrambled. Again, try to figure out the meaning and/or identify the missing words.

> Brad Pitt told Esquire magazine that he and Angelina Jolie will not be married until the right to marry is given to gays and plortes. Pitt, who graced the cover of the magazine's October issue, says, "Angie and I will consider tying the tonk when everyone else in the country who wants to be married is able."

How does Lexxica measure vocabulary size?

Lexxica's online V-Check test and V-Flash level check employ similar regimented data collection and statistical processes to determine the mathematical probability of any semantic or lexical item being recognized by a member of a particular population group.

The collective responses of a population group are compiled to establish the aggregate difficulty factor for each semantic item. The result is what we call a Recognition Ogive that is unique to its population group.

When taking a Lexxica test or level check, an individual is presented with semantic items (words, terms, expressions, polywords, idioms, constructs, signs, images, etc.) selected from points along a Recognition Ogive belonging to the user's population group. Lexxica's IRT based Computer Adaptive Test, quickly determines an accurate measure of user ability along the user's Recognition Ogive. An essential and proprietary element of the process is our inclusion of false semantic items to control for guessing. False semantic items are introduced in accordance with standard precepts of Signal Detection Theory.

Lexical item difficulty data is systematically collected with each new Lexxica test. Lexxica's processes are adept at assessing not only Base Word recognition but also Base Word depth of knowledge. Over time, our understanding of how the brain treats and processes all forms of semantic items will be greatly advanced.

What have we learned thus far?

Several important observations have emerged.

1) Word difficulty can differ greatly from one population group to another. Accurate measurement of vocabulary size requires a process that is capable of independently measuring word difficulty for each specific population group.

2) Even within the same population group, word frequency is not a good predictor of word difficulty. While there is some overall correlation (0.6) between word

frequency and word difficulty the standard error from one word to the next and from one person to the next are so high that any approach that attempts to extrapolate vocabulary size from recognition of a subset of frequent words will produce statistically unreliable findings.

3) Based on 10,000+ surveys, even Japanese students having a large vocabulary (3000+ words) are often missing 400 or more words from among the first 2000 most important and high-frequency words. This suggests that direct vocabulary instruction can quickly provide the missing words and fill the knowledge gaps – provided, that is, that the learner's missing words are accurately identified.

4) The Recognition Ogive established for the Japanese population permits us to almost instantly identify the specific important and high frequency words that each learner does not already know. This enables the generation of individualized learning programs based on each learner's precise needs analysis. The popularity of mobile phones and PCs has made it possible to efficiently test and instruct vocabulary in ways that were previously inconceivable.

5) We believe that the brain stores and processes lemmas having similar difficulty factors as forms of the same word, and that the brain stores and processes lemmas having different difficulty factors as different words. Our hypothesis is based on findings such as the difficulty factors for the words 'accept' 'accepts' 'accepting' and 'unacceptable' among Japanese respondents. Indications are that 'accept' 'accepts' and 'accepting' have almost exactly the same statistical difficulty factor. But the word 'unacceptable' has a significantly higher difficulty factor. Under the word family approach, 'unacceptable' would be counted together as one with 'accept'. Under our hypothesis, we conclude that 'unacceptable' is a different word and not a form of the word 'accept.'

6) A Recognition Ogive established among native speakers (including top performers and experts where a special purpose vocabulary is concerned) provides an exciting new measure of lexical item importance. Corpus item frequency counts tell us something about how words have been used in the past. Statistical item difficulty among a population tells us about how words are known and used by people today. In determining the most important general vocabulary words for students to learn, it is highly instructive to understand not only frequency but also the rank order of word difficulty among native speakers. Likewise, in selecting the most important special terms to a career or personal interest, it is immensely useful to know the rank order of word difficulty from beginners up to top performers and experts.

Background and Research

Lexxica has developed a comprehensive system for individualized lexical knowledge assessments and the generation of individualized courses of language study. Most of the references in this report describe the assessment and instruction of English, however, the reader should bear in mind that all of the processes described apply equally to the assessment and instruction of languages (and semantic systems) other than the English language. Furthermore it is the express intention of Lexxica to apply its systems to a wide variety of languages, and language sub-domains, irrespective of whether they are a native language or a non-native language of the learner.

Until recently, vocabulary learning was seen as peripheral to language acquisition, both theoretically and practically. Linguistic theory assigned word learning to a simple functional-associative model which of course could not accommodate syntax, and applied language researchers and teachers largely concurred with this view in an effort to be aligned with proper theories, and also in the knowledge that vocabulary was anyway too vast a quantity for direct instruction (but fortunately could be picked up more or less by itself).

With the grammar-translation method, and its focus on the syntax of the sentence, it was thought that once the students learned the grammar of the sentences, they would be able to slot in vocabulary and therefore generate language. The advent of the Audiolingual method, based on habit-formation, was much the same regarding vocabulary. Words were taught only within the structures that were the main focus. Since then, subsequent research has often attempted to account for second language acquisition (“SLA”) by looking at grammatical features in such areas as the developmental sequence (Cancino, Rosansky, & Schumann, 1978; Pienemann, 1989), the role of input (Loschky, 1994; Shook, 1994; White, Spada, Lightbown, & Ranta, 1991), and instruction (Dulay, & Burt, 1973; Ellis, 1992; Sharwood Smith, 1981; VanPatten, & Cadierno, 1993). From the publication of Corder's seminal paper in 1967 to Larsen-Freeman writing in 1991 on SLA research, the study of grammar and its acquisition has almost become synonymous with SLA.

Much of what was believed has now been reversed. Theoretically, it appears likely that language acquisition begins with word learning rather than syntax triggering, with words gradually "grammaticalized" through experience on a largely associative basis. Practically, studies throughout the 1980s and 1990s showed that vocabulary skill and knowledge are the precondition for most other language abilities and, in addition, the main source of variance in the final state of such abilities. It now seems clear that vocabulary acquisition does not happen by itself to any satisfactory degree, particularly as needed for first language literacy or a second language generally.

Over the years, a relatively small group of scholars has worked consistently to consider the needs of learners from a predominantly lexical perspective. Many of the questions they asked, and the results they found are still relevant today. These questions included how many words a student needed to know, how these words should be sequenced, and what the student needed to know about these words.

One of the first debates centered around the number of words that a student needed to know. This necessarily led to defining what a word is, and what it means to know a word. While this research primarily focused on first language acquisition, there are obvious implications for SLA as well. The central argument was whether it would be possible to increase a learner's vocabulary by the direct instruction of words and their meaning. If estimates of native speakers vocabulary were large, explicit instruction would not be feasible, and early research seemed to indicate that this was the case. Studies cited in D'Anna, Zechmeister, & Hall, (1991) suggested a recognition vocabulary of 155,736 words (Seashore & Eckerson, 1940) and over 200,000 words (Hartman, 1941) but both studies suffered from methodological problems in defining what a word is.

Nagy and Anderson (1984) used six semantic categories to organize lexis from a corpus of high school English and found that students were exposed to 45,000 base words and 88,500 word families. They suggested that teaching children "words one by one, ten by ten, or even hundred by hundred would appear to be an exercise in futility" (p. 328), and that teachers should concentrate on teaching skills and strategies for independent word learning. Later research by Goulden, Nation, & Read, (1990) questioned whether native speakers actually knew these words. By designing tests based on the frequencies of the words, the researchers determined that native speakers' vocabulary averages 17,200 words. This number suggests that the learning burden is not as insurmountable as previously suggested. Other research by D'Anna, et al. (1991) found a similar result of 16,785 words.

Hazenberg & Hulstijn, (1996) found that native Dutch speakers had a vocabulary of 18,807 words but they also looked at the vocabulary of non-native students writing a Dutch university entrance exam and concluded that these students needed a minimum of 10,000 base words for entry into university. Laufer (1989) compared vocabulary size and reading comprehension scores and found that a recognition vocabulary of at least 3,000 words was a threshold for being able to read unsimplified texts. While in no way negating Nagy and Anderson's argument that learning vocabulary from reading is important, there is sufficient evidence that teaching at least some of the words explicitly can have a meaningful effect on the students' vocabulary.

While an assessment of vocabulary size provided part of the picture, other researchers looked at which words ESL students needed and how they should be sequenced. In the 1930's through to the 1950's, a few researchers (Ogden, 1930; Richards, 1943; Thorndike, & Lorge, 1938; West, 1953) ranked vocabulary, using criterion mostly based on frequency and coverage. Ogden's 850 basic words and West's 2000 word general service list sought to provide a way to assist the learners in acquiring a sufficient vocabulary to overcome what Coady (1993) would later refer to as "the paradox of learning words through context", whereby students must have a command of enough words to read in the first place. With the 2,000 high-frequency words accounting for 81 percent of the running words in a text (Nation, 2001), students who have mastered this list are better prepared to handle the demands of reading.

However, research by Laufer (1989) clearly shows that even this amount may not be sufficient for academic study in an L2 environment or reading unsimplified texts.
Another important issue involved the depth of knowledge necessary to understand the various dimensions of a meaningful and full representation of a given word. Depth of word knowledge has been categorized along a continuum from receptive to productive, into four categories consisting of form, position, function and meaning (Nation, 1990), or into comprehension processes (Quian, 1999) including pronunciation and spelling, morphological properties, syntactic properties, meaning, register, and word frequency. Many techniques have been suggested to increase vocabulary knowledge (Bauer & Nation, 1993; Crow & Quigley, 1985; Hafiz & Tudor, 1990; Joe, 1995; Nation, 1990, 1994a, 1994b; Nattinger, 1988; Williams, 1986; Wodinsky & Nation, 1988), varying on the explicitness of the presentation from word list memorization techniques (Crow & Quigley, 1985) to learning through communicative interaction (Joe, 1995). The need to expand vocabulary learning in line with overall linguistic development has received considerable attention (Carter & McCarthy, 1988; Chall, 1987; Nation, 1990, 1994a, 1994b; Parry, 1993), but until now there hasn't been a technologically feasible way to achieve this expansion. Emerging technologies in communication and personal computers are ideally suited to support and advance understanding of vocabulary such that an efficient, personalized learning experience can be provided.

According to Brown, (1995) an essential component of any pedagogical program is a needs analysis. Before designing and presenting materials, it is imperative to gather "information to find out how much the students already know and what they still need to learn" (p.35). In a vocabulary program, the first requirement is to identify what words the students need to learn through the analysis of corpora. The second procedure is to test the words to find out how many of the words the student already knows.

Since the pioneering work of George Kingsley Zipf and E. L. Thorndike, the statistics analyses of large collections of texts have helped to determine some of the more valuable properties of usage. One such field of study has to do with the relationship between the rank of a word, the frequency to which it occurs in text, and the cumulative coverage of the text. The most common word in English, *the*, occurs about 7 times in every 100 words of text. About a quarter of all the words in a text will be one of the 10 most common words. As words become less frequent, their contribution to the coverage of the text decreases. While the 100 most common words account for about half of all the words in a text, the next 100 only account for 7 percent, bringing the coverage up to 57 percent of the running words in a text. Nation (1990) summarizes Carrol, Davies, and Richman research on frequency counts in the Brown corpus in the Figure 1 below. Column one represents the cumulative number of words starting from the highest frequency. There are 86,741 different words in the Brown corpus. The second column shows the percentage of words in the corpus that the words account for. For example, the 10 most frequent words account for 23.7 percent of all the words in the corpus.

Figure 1

Different words	Percentage of running words
86,741	100
43,831	99
5,000	89.4
3,000	85.2
2,000	81.3
100	49
10	23.7

While these figures differ slightly from corpus to corpus, the general trend is consistent. After about 2000 words, lower frequency words contribute little to the coverage. Learners with less than 2000 words would have great difficulty comprehending natural text, as approximately one out of every five words is unknown. Learners with 2000 to 3000 word vocabularies would still struggle with the text., are percent general,families ed for 95 percent coverage would be much lower

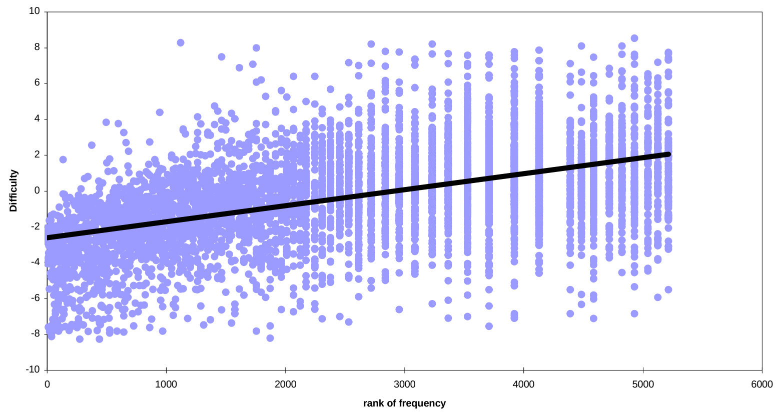
At this stage, however, it is possible to extract lexical units that are common and frequent to a given genre of text, by comparing their frequency in the genre to their expected frequency in general text. By this process, we can identify vocabulary for special purposes. For example, in the 100 million words British National Corpus, the word *nocturnal* appears twice per million words. In a book about wildlife, we would expect to see it more frequently than that. This deviation, clustered with similar lexical deviations, would identify the text as being different from the general text. Alternatively, by analyzing genre specific text, we can identity the specialized vocabulary. This process has been used to compile academic word lists (Coxhead, 2000, Xue & Nation, 1984). One of our goals is to identify and compile lists of words and multiword lexical units for a number of fields.

Traditionally, after having identified which words were necessary for the learners to learn in order to comprehend a written text, researchers such as Thorndike and Lorge (1944), and West, (1955) evaluated the words for usability and generalizability in order to compile a list for teaching. Until now, however, given the large number of words, and the problems with test equating and item indices under classical test theory, it has been practically impossible to find out which words the students knew. Scores on different tests by different groups could not be compared. Word frequency is used as a substitute for word difficulty.
The advent of Item Response Theory (“IRT”) in the late 1950s and 60s brought with it the benefits of large scale testing, and the ability to assign a score to the difficulty of an item regardless of the group who took the test. The System uses an unique IRT model to estimate word difficulty from large scale vocabulary testing, and applies the findings to generate both ability estimates for the person and specific sequences of target vocabulary for learning.

dItem response theory is a probabilistic model that attempts to explain the response of a person to an item. The probabilities of a given response can be expressed mathematically through a number of different formulas, depending upon the situation. In its simplest form, item response theory posits that the probability of a random person *j* with ability θ_j answering a random item *i* with difficulty b_i correctly is conditioned upon the ability of the person and the difficulty of the item. In other words, if a person has a high ability in a particular field, he or she will probably get an easy item correct. Conversely, if a person has a low ability and the item is difficult, he or she will probably get the item wrong. When we analyze item responses, we are trying to answer the question, "what is the probability of a person with a given ability responding correctly to an item with a given difficulty?"

With large-scale testing of our wordlists, we have been able to compare the measure difficulty of the word with a mathematical manifestation of the rank of the word. This can be seen in Figure 2 below. The horizontal axis represents the ranking of the frequency of the words. The data are arranged in ascending order, with the highest frequency words on the left. For this particular manifestation, the vertical axis shows the difficulty index as calculated from 4,217 Yes\No tests on 6000 words. The data are arranged in increasing difficulty with the easiest items at the bottom and the more difficult items at the top. The data shows the relationship between frequency and difficulty as represented by the regression line. It also shows that there are many words of low frequency that are well recognized, and there are many high frequency words that may not be known.

<u>Figure 2</u>



Yes/No tests, also known as lexical decision tasks, ask learners to identify known words from a list of real and non-words, or pseudo-words. While these types of tests are not common in most language classrooms, they have a long history in research in the field of psycholinguistics, where they have played an important role in our understanding of how the mental lexicon works. These tests are often analyzed using a branch of Decision Theory known as Signal Detection Theory ("SDT"), which compares the learner's responses to the real words and non-words, and determines the probability of a correct decision as well as the degree of accuracy to which the learner makes the decision. With the increasing availability of computer adaptive testing, these methods are now making the jump from the research lab to the digital classroom.

Unlike convention pencil and paper tests where the reliability and accuracy of the test can only be established through the statistical analysis of the responses after the test has been taken, CAT predetermines the level of accuracy, then in an interactive manner administers items, based upon the response pattern of the test taker, until the desired level of accuracy has been achieved. Since the test is constantly zeroing in on a respondent's level based on their correct or incorrect responses, a far fewer number of questions are needed to accurately estimate their level.

The accuracy of a measure is associated with the Standard Error of Measurement ("SEM"). With conventional pencil and paper tests, the SEM is derived from the Standard Deviation ("SD") and reliability of the test as shown in Formula 1 below.

$$SEM = SD\sqrt{1-r} \qquad (1)$$

where *SEM* is the standard error of measure
SD is the standard deviation
r is the reliability of the test

With IRT, the standard error of the estimate, a statistic related to SEM, is derived from the amount of information that each item contributes to the test results. Formula 2 shows the information function for the estimate based on a test, and Formula 3 illustrates the relationship with the standard error of the estimate.

$$I(\theta) = \sum_{i=1}^{n} \frac{\left[P_i'(\theta)\right]^2}{P_i(\theta)Q_i(\theta)} \qquad (2)$$

where $I(\theta)$ is the information provided by a test of items 1 to *n*
$P_i'(\theta)$ is the derivative of $P_i(\theta)$.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad (3)$$

where $SE(\theta)$ is the standard error of the estimate

In a CAT, the respondent is presented with the first item, usually drawn from a pool of items very close to the population mean. Depending on how the test taker responds, the next item will be drawn from approximately one standard deviation from the mean. This will continue until there is at least one item answered correctly, or in the case of a Yes/No test, one real word is identified as being known, and one item answered incorrectly, or one real words identified as unknown. At this point, a maximum likelihood estimate of the test-taker is calculated using the derivative of the likelihood function, as well as the test information function and standard error shown above.

Each next item is selected to give the maximum amount of information at the estimate of the ability. Then the maximum likelihood, test information, and standard error of the estimate are calculated again. This process is repeated until the desired level of accuracy is achieved. This process depends on the responses of the test taker, therefore, the amount of time necessary to take the test is variable. However, because each item is selected to maximize the information and minimize the error based on an individuals responses, these tests are always more efficient than conventional pencil and paper tests or non-interactive computer tests.

Much research in the area of second language vocabulary acquisition has focused or depended on estimates of a learner's overall vocabulary size (i.e., breadth of vocabulary). Tests such as Nation's (2001) 'Vocabulary Levels Test' attempt to measure respondent's passive recognition of vocabulary words at different frequency bands for purposes such as measuring group gains, program evaluation, or student placement. While useful, such tests have a number of limitations, including an inability to assess how well particular words are known (Read, 1988). More recent work (Nassaji, 2004, Waring, 2002, Vermeer, 2001, Paribaht & Wesche, 1993, Wesche & Paribaht, 1996), has begun to explore how to assess a learners' level of familiarity with a given word.

In general, knowledge of a given lexical item is considered to exist on a continuum of less knowledge to more knowledge (see Figure 4), from a receptive understanding of the item at the beginning stages to a more productive understanding at later stages of learning. In other words, early stages of vocabulary knowledge might include the receptive ability of being able to recognize a word in a written sentence or stream of speech, while later stages might include the ability to use the word productively in a written or spoken sentence.

Over the years a variety of depth of knowledge scales based on student self-assessment-type questionnaires have been developed. These include Eichholz and Barbe's test of word knowledge (1961), D'Anna and Zechmeister's vocabulary knowledge scale (1991) and Zimmerman's 4-point vocabulary knowledge scale (1997).

Lexxica's assessment of vocabulary knowledge draws several ideas from the above models in order to provide a fast and efficient means of assessing certain aspects of a respondent's depth of vocabulary knowledge utilizing an interactive computer interface. In order to make the system's online test as efficient as possible, no depth of knowledge questions are asked until after the Yes/No section of the test is complete and the system has been able to determine the approximate number of words the respondent knows. Once this has been established, a small number of depth of knowledge questions will be asked, at the respondent's estimated level of difficulty and next at progressively lower difficulty levels. The reason for testing at lower levels is that respondents' depth of knowledge of words located toward the high end of the respondent's level of difficulty will most likely be quite shallow. Deeper understanding is to be expected for easier words. The system seeks to generate information about a respondent's depth of knowledge at different levels of difficulty in order to best determine a more useful and effective individualized course of study.

For non-native language knowledge assessments, the system is capable of testing respondents on certain words that have been identified by the respondent as being known, in order to ascertain which of these items are *false-friends* (i.e., words from the respondent's native mother-tongue that are spelled or sound like words in the non-native language being tested but whose usage or meaning in the native language is actually very different), and which are genuinely known.

Once the system has obtained the respondent's ability estimate based on the test results, the system can convert the score into an estimate of the number of words the respondent knows through use of its regression formula. By converting ability estimates into the number of words known, respondents and their teachers can receive a useful absolute assessment of language knowledge. Not only can the assessment score be used to accurately gauge a respondent's learning progress over periods of time, it can also provide a more meaningful way to interpret respondent test results, and it can be used to create, select or assign ability appropriate graded reading material at any level of ability.

In developing individualized courses of vocabulary study, one likely approach would be to prioritize words as follows: The first group of words to be presented for study would be important and highly frequent general vocabulary words for which a learner has indicated a low depth of knowledge. In other words, common lexical items that a respondent thinks they know, but of which they have little or incorrect knowledge as revealed by the test. The next group of words to be presented would be important high frequency general vocabulary words at a slightly higher difficulty level. These words will be presented in order of importance, as will all words within their specific sub-domain. Where possible (not all learners have a special field of interest), the next group of words to be presented would be drawn from specialized words appropriate to a learner's professional field or area of special interest at or near the learner's assessed level of ability, and for which indications are that little depth of knowledge is possessed. The fourth group of words to be presented would be specialized words appropriate to a learner's professional field or area of interest that are above a learner's assessed ability level. The fifth group of words to be presented would be important low frequency general vocabulary words slightly above a learners' assessed vocabulary ability. The sixth group of words to be presented for study would be important low frequency general vocabulary words that are well above a learner's assessed vocabulary size and at a higher level of difficulty.

Acknowledgement

Our efforts are logical extensions of the excellent research and conclusions of the many pioneering lexical researchers who have come before us. We welcome constructive comments and contributions from researchers that may help advance our efforts. We look forward to supporting the application of Lexxica processes and technology to research in linguistics, educational psychology and cognitive neuroscience. We are particularly interested in work done in the areas of; (i) lexical recognition, production and storage in the brain; (ii) relationships between forms of input and adult acquisition and retention; (iii) how children develop and structure lexical knowledge; (iv) the nature and scope of cross-linguistic universals of lexical structure and acquisition; (v) the effects of atypical degenerative or traumatic brain conditions on lexical and cognitive processing; and, (vi) the relationship between lexical unit knowledge and general cognition. We invite researchers seeking support or grants to submit a one-page abstract of their research proposal to research@lexxica.com.

References

Bauer, L. & Nation, I. S. P. (1993). Word families. International Journal of Lexicography, 6, 253-279.

Cancino, H., Rosansky, E. J., & Schumann, J. H. (1978). The acquisition of English negatives and interrogatives by native Spanish speakers. In E. Hatch (Ed.), Second language acquisition (pp. 207-230). Rowley, MA: Newbury House.

Carter, R. & McCarthy, M. (Eds.). (1988). Vocabulary and language teaching. London: Longman.

Chall, J. S. (1987). Two vocabularies for reading: Recognition and meaning. In M. G. McKeown, & M. E. Curtis (Eds.), The nature of vocabulary acquisition (pp. 7-17.). Hillsdale, NJ: Lawrence Erlbaum.

Coady, J. (1993). Research on ESL/EFL Vocabulary Acquisition: Putting it in context. In T. Huckin, M. Haynes, & J. Coady (Eds.), Second language reading and vocabulary learning (pp. 3-23). Norwood, NJ: Ablex Publishing.

Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics, 5, 161-170.

Crow, J. T., & Quigley, J. R. (1985). A semantic field approach to passive vocabulary acquisition for reading comprehension, TESOL Quarterly, 19, 497-513.

D'Anna, C. A., Zechmeister, E. B. & Hall, J. W. (1991). toward a meaningful definition of vocabulary size. Journal of Reading Behavior, 23, 109-122.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? Language Learning 23,(2), 245-258.

Goulden, R. Nation, I. S. P., & Read, J. (1990). How large can a receptive vocabulary be? Applied Linguistics, 11,(4) 341-363.

Hafiz, F. M. & Tudor, I. (1990). Graded readers as an input medium in L2 learning. System, 18, 31-42.

Hazenberg, S. & Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. Applied Linguistics, 17, 145-163.

Joe, A. (1995). Test-Based tasks and incidental vocabulary learning. Second Language Research, 11, 149-158.

Larsen-Freeman, D. (1991). Second language acquisition research: Staking out the territory. TESOL QUARTERLY, 25, 315-350.
Laufer, B. (1989). How much lexis is necessary for reading comprehension. In Vocabulary and Applied Linguistics (pp. 126-132).
Loschky, L. (1994). Comprehensible input and second language acquisition: what is the relationship? SSLA, 16, 303-323.
Nagy, W. E. & Anderson, R. C. (1984). How many words are there in printed school English? Reading Research Quarterly, 19, 304-330.
Nation, I. S. P. (1990). Teaching and learning vocabulary. Boston: Heinle & Heinle Publishers.
Nation, I. S. P. (1994). Teaching listening and speaking. Tokyo: Temple University Japan
Nation, I. S. P. (Ed.). (1994). New ways in teaching vocabulary. Alexandria, Virginia: TESOL
Nattinger, J. (1988). Some current trends in vocabulary teaching. In R. Carter, & M. McCarthy (Eds.), Vocabulary and language teaching. London: Longman
Ogden, C. K. (1930). Basic English: A general introduction. London: Kegan Paul, Trench and Trubner.
Parry, K. (19). Too many words: Learning the vocabulary of an academic subject. In T. Huckin, M. Haynes, & J. Coady (Eds.), Second language reading and vocabulary learning (pp. 109-127). Norwood, NJ: Ablex Publishing.
Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypotheses. Applied Linguistics, 10, 52-79.
Qian, D.D. (1999). Assessing the role of depth and breadth of vocabulary knowledge in reading comprehension. Canadian Modern Language Review, 56, 282-308.
Richards, I. A. (1943). Basic English and its uses. London: Kegan Paul.
Sharwood Smith, M. (1981). Consciousness raising and the second language learner. Applied Linguistics, 2, 159-168.
Shook, D.J. (1994). FL/L2 reading, grammatical information, and the input-to-intake phenomenon. Applied Language Learning, 5(2), 57-93.
Thorndike, E. L. & Lorge, I. (1938). A semantic count of English words. New York: Columbia University Press.
VanPatten, B. & Cadierno, T. (1993). Explicit instruction and input processing. Studies in Second Language Acquisition, 15, 225-243.
Webster's third new international dictionary. (1961). Springfield, MA: Merrian-Webster Inc.
West, M. (1953). A General Service List of English Words. London: Longmans, Green and Co. Ltd.
White, L., Spada, N., Lightbown, P. & Ranta, L. (1991). Input enhancement and L2 question formation. Applied Linguistics, 12, 416-432.
Williams, R. (1986). Teaching vocabualry recognition strategies in ESP reading. The ESP Journal, 4, 121-131.
Wodinshy, M. & Nation, I. S. P. (1988). Learning from graded readers. Reading in a foreign language, 5, 155-161.
Bauer, L. & Nation, I. S. P. (1993). Word families. International Journal of Lexicography, 6, 253-279.
Brown, J. D. (1995). The elements of language curriculum: A systematic approach to program development. Boston: Heinle & Heinle Publishers.
Cancino, H., Rosansky, E. J., & Schumann, J. H. (1978). The acquisition of English negatives and interrogatives by native Spanish speakers. In E. Hatch (Ed.), Second language acquisition (pp. 207-230). Rowley, MA: Newbury House.
Carter, R. & McCarthy, M. (Eds.). (1988). Vocabulary and language teaching. London: Longman.
Chall, J. S. (1987). Two vocabularies for reading: Recognition and meaning. In M. G. McKeown, & M. E. Curtis (Eds.), The nature of vocabulary acquisition (pp. 7-17.). Hillsdale, NJ: Lawrence Erlbaum.

Coady, J. (1993). Research on ESL/EFL Vocabulary Acquisition: Putting it in context. In T. Huckin, M. Haynes, & J. Coady (Eds.), Second language reading and vocabulary learning (pp. 3-23). Norwood, NJ: Ablex Publishing.
Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics, 5, 161-170.
Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34, 213-238.
Crow, J. T., & Quigley, J. R. (1985). A semantic field approach to passive vocabulary acquisition for reading comprehension, TESOL Quarterly, 19, 497-513.
D'Anna, C. A., Zechmeister, E. B. & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. Journal of Reading Behavior, 23, 109-122.
Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? Language Learning, 23, 245-258.
Eichholz, G., & Barbe, R. (1961). An experiment in vocabulary development. Educational Research Bulletin, 40, 1-7.
Ellis, R. (1992). Grammar teaching - practice or consciousness-raising?. In R. Ellis (Ed.) Second language acquisition and language pedagogy (pp.). Clevedon, Avon, Multilingual Matters.
Goulden, R. Nation, I. S. P., & Read, J. (1990). How large can a receptive vocabulary be? Applied Linguistics, 11, 341-363.
Hafiz, F. M. & Tudor, I. (1990). Graded readers as an input medium in L2 learning. System, 18, 31-42.
Hazenberg, S. & Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. Applied Linguistics, 17, 145-163.
Joe, A. (1995). Test-Based tasks and incidental vocabulary learning. Second Language Research, 11, 149-158.
Larsen-Freeman, D. (1991). Second language acquisition research: Staking out the territory. TESOL Quarterly, 25, 315-350.
Laufer, B. (1989). What percentage of text-lexis is essential for comprehension?. In C. Lauren & M. Nordman (Eds.) Special language: From humans thinking to thinking machines (pp. 316-323). Clevedon, England, Multilingual Matters.
Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Béjoint & P. Arnaud (Eds.), Vocabulary and applied linguistics (pp. 126-132). London: Macmillan.
Loschky, L. (1994). Comprehensible input and second language acquisition: what is the relationship? SSLA, 16, 303-323.
Nagy, W. E. & Anderson, R. C. (1984). How many words are there in printed school English? Reading Research Quarterly, 19, 304-330.
Nassaji, H. (2004). Input modaility and remembering name-referent associations in vocabulary learnings. Candian Journal of Applied Linguistics 7, 39-56.
Nation, I. S. P. (1990). Teaching and learning vocabulary. Boston: Heinle & Heinle Publishers.
Nation, I. S. P. (1991). Learning vocabulary in another language. Cambridge: Cambridge University Press.
Nation, I. S. P. (1994a). Teaching listening and speaking. Tokyo: Temple University Japan
Nation, I. S. P. (Ed.). (1994b). New ways in teaching vocabulary. Alexandria, Virginia: TESOL
Nattinger, J. (1988). Some current trends in vocabulary teaching. In R. Carter, & M. McCarthy (Eds.), Vocabulary and language teaching. London: Longman
Ogden, C. K. (1930). Basic English: A general introduction. London: Kegan Paul, Trench and Trubner.
Paribakht, T. S. and Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. TESL Canada Journal. 11, 9-29.
Parry, K. (1993). Too many words: Learning the vocabulary of an academic subject. In T. Huckin, M. Haynes, & J. Coady (Eds.), Second language reading and vocabulary learning (pp. 109-127). Norwood, NJ: Ablex Publishing.

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypotheses. Applied Linguistics, 10, 52-79.
Qian, D.D. (1999). Assessing the role of depth and breadth of vocabulary knowledge in reading comprehension. Canadian Modern Language Review, 56, 282-308.
Rasch, G. (1993). Probabilistic models for some intelligence and attainment tests, with a forward and afterword by B. D. Wright. Chicago: Mesa Press. (1960, Danish Institute for Education Research)
Read, J. (1988). Measuring the vocabulary knowledge of second language learners. RELC Journal, 19, 12-25.
Richards, I. A. (1943). Basic English and its uses. London: Kegan Paul.
Sharwood Smith, M. (1981). Consciousness raising and the second language learner. Applied Linguistics, 2, 159-168.
Shook, D.J. (1994). FL/L2 reading, grammatical information, and the input-to-intake phenomenon. Applied Language Learning, 5, 57-93.
Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriatemess. Applied Psychological Measurement, 11, 335-370.
Thorndike, E. L. & Lorge, I. (1938). A semantic count of English words. New York: Columbia University Press.
Thorndike, E. L., & Lorge, I. (1944). The teacher's Word Book of 30,000 Words. New York: Teachers College, Columbia University.
VanPatten, B. & Cadierno, T. (1993). Explicit instruction and input processing. Studies in Second Language Acquisition, 15, 225-243.
Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. Applied Psycholinguistics, 22, 217-235.