Vocabulary Testing: Perspectives and Prospects

Daniel DUNKLEY

Vocabulary learning occupies a special place in all types of language learning. We acquire competencies at widely differing speeds. In our native language, although we acquire proficiency in the phonological system very early, and our grammatical skill is fairly complete after about 15 years, vocabulary learning occupies us throughout our period of full-time education, and is never completely finished. In fact whereas all educated adults have a common level of pronunciation and grammar, they share only a core body of vocabulary; a considerable part of our vocabulary is dependent on our life experiences, both professional and private.

In second language learning, too, vocabulary is a special case. The phonological system of a foreign language can be understood fairly quickly, since there is a limited number of sounds the human speech production system is capable of, and thus there is a considerable area of overlap between the sounds of any two languages. However, the irony with phonology is that in spite of this similarity, for most learners phonology is an insuperable barrier, and one is almost bound to be recognized by native speakers as foreign in the sounds one makes. Grammar too, can normally be understood in a year or two by an adult, and with practice it can be mastered practically. But vocabulary learning is a slow and laborious process, involving considerable trial and error, and repeated forgetting and re-learning.

Any assessment of spoken or written competence in a foreign language involves vocabulary testing. This paper will examine the nature of vocabulary, the process of vocabulary acquisition, the state of research into vocabulary assessment, and future prospects for improvements in vocabulary testing.

Vocabulary fundamentals

The layman would be surprised to be asked the question "What is a word? The average language learner, too, is familiar with bilingual vocabulary lists, and can probably state roughly how many words he or she knows in the language in question. However, the definition of a word is a thorny one for many specialists, particularly lexicographers and testers. To show the need for refinement of the blunt idea of a word, we could start by counting the words in a paragraph. Here we notice that some words occur frequently especially *a, the* and *that*, while others occur only once. This distinction leads to the idea of a *token*, the individual word, and *type* the different word forms. Thus in the brief text "*He has a brother and a sister*." there are seven tokens but only six types. It is well known that as a speaker of a language, whether native or second, matures, the ratio of types to tokens increases. A child might say *My brother has a computer and my sister has a computer*. (11 tokens 7 types) A more mature speaker might say *Of my siblings, my brother and sister both have computers*. which contains 11 tokens and 11 types.

The next point is that when we speak of vocabulary size we need to decide whether we should count words like *a the to for* and so on as words. There is clearly a continuum of meaning. *A* and *the*, the articles, while being an integral part of many languages (for example English, French and German) are absent from many others (Russian, Japanese). Thus they are virtually dispensable, and in fact many speakers of English, French and German as second languages around the world communicate quite effectively without using them. At the other end of the continuum are nouns, which if removed from a sentence make it meaningless. Somewhere between these two extremes come prepositions and adjectives. Words at the low end of this hierarchy of meaning are sometimes called *function words* and at the high end *content words*.

Words in most language are inflected in several ways: using suffixes, prefixes or by changes in a middle vowel. Thus we find *sing sings*, (Eng) *plyvat' poplyvat'* (Russian) or *sing-sang* (Eng, Ger) or by; a combination of two or more methods of inflection *Land-Länder* (German) *plyvat' poplyval* (Russian). The collective term for a base form and one or more inflected forms is a *lemma*. Thus, for example the group *sing sings sang sung* is one lemma.

While there are several inflected forms of words – in English essentially nouns pronouns and verbs, but in many other languages also articles and adjectives – there are also related forms of a word with distinct meanings. Thus the noun *board* has the inflected form *boards*, then the verb forms *board boards* and *boarded*, and in addition several other meanings as a verb: to board a ship to board at a school *to board up* a window. Even as a noun it has three distinct meanings: *a piece of wood* and *a committee* and *accommodation* (*room and board*). This large collection of related words is referred to as a *word family*.

The above distinctions are important when we come to estimate vocabulary size. Essentially a count of vocabulary size should be a count of knowledge of word families, not of word forms. For example if we count *run runs ran and running* as four words rather than as one word family *run*, then clearly we will credit a speaker with three times the actual number of words. This is one reason why estimates of vocabulary size vary so widely. One recent survey noted that an English native speaker's vocabulary been put at anything from 18,000 to

25,000 words. (Quinion 2010)

In addition to the question of defining a word, it is clear that we need to be aware not just of individual words but also of phrases or idioms, sometimes called *multi-word lexical items*. One example of a study of this field is that of Nattinger and DeCarrico (1992). They adopted the terminology *lexical phrase*, and proposed four types: the *polyword* (for the most part) the *institutional expression* (once upon a time) phrasal constraints (a … ago) and finally *sentence builders* (Not only … but also …).

Although multi-word lexical items are an important part of our vocabulary, we find that most vocabulary tests consist of single word items. There are several reasons for this. It is one thing to identify multi-word lexical items but quite another to assess knowledge of them. Especially when using computer-based texts, it is much easier to count single word items than multi-word items. Furthermore, they are very slippery, in that they vary in form and are not a finite easily identified group. A final reason for their absence from most tests is their comparative rarity, as Read (2000, 24) points out "If frequency in the language is an important criterion for choosing lexical items for a selective vocabulary test, only a small number of multi-word items may qualify on that basis."

What do we mean when we say "I know that word?" Before we test something we must define it, or to put it in more formal terms, we must define a construct before operationalizing it. The simple understanding of knowing is that we can remember words. But what do we mean by that? Is it just that we can produce it at the right time? When we consider everyday language use, we realize that native speakers' knowledge is not the same for every word. Whereas some words can be spontaneously produced, others can be recalled only with effort, and still others can not be produced at all, but only recognized when heard or read.

In fact, rather than simply having the word in one's memory, knowledge of an item involves a variety of quite separate skills. According to one well known article (Richards 1976) there are seven distinct sub-skills in knowing a word. For example, there is knowing in the sense of knowing the underlying form and its derived forms; also in the sense of knowing the limitations of the use of the word. An example of the latter knowledge is to say when we know the word *station* correctly we will not stay *train stop* for *train station* nor will we say *bus station* when we mean *bus stop*. A more complex picture of vocabulary knowledge is given by Nation (1990) who divides it into productive and receptive skills. In fact these two areas are unbalanced; production involves a greater knowledge of a word than reception. Nation's scheme divides knowledge of a word into knowledge of form, position, function and meaning.

Whereas testers are always concerned to make a definition of vocabulary knowledge fit their particular needs, applied linguists are not constrained in this way. So a wide-ranging characterization of vocabulary knowledge has been proposed by Chapelle (Chapelle 1994, Chapelle and Read 2001). Her scheme combines two notions: that of knowledge and use. Accordingly there are three main strands to her explanation: the context of vocabulary use, (both linguistic and pragmatic), vocabulary knowledge, similar to the previous definition and finally metacognitive strategies. The most common skill in the latter case is avoidance, by which a speaker or writer sidesteps an unknown word.

In summary, we have seen how vocabulary knowledge involves both many different kinds of knowledge and also a range of skills, and includes the use of words in context. It is an awareness of this complexity which leads the language tester to make decisions on the strategy to be applied when testing vocabulary knowledge.

Learning vocabulary

While the language tester needs to be equipped with a practically useful definition of vocabulary knowledge, he or she is greatly helped by an understanding of vocabulary acquisition. Of course, most people have learned a second language and have memories of how they learned vocabulary. The average English native speaker's first experience of language learning was probably French or Spanish, so the strategy of looking for similar L1 words probably comes to mind first. However, thinking globally, for most language students learning a foreign language is a very different experience because there is so little in common with one's native language. In recent years there has been an effort to understand how students learn words, (see Laufer 1998) particularly focusing on two questions. The first one is leaning facility: what makes a work difficult to learn? The second one is retention: what is the best means of ensuring that a word is remembered?

It has been found that different classes of words have different levels of difficulty. For example, nouns are easier to learn than adjectives. Then, idioms and phrases are more difficult than adjectives. Finally the word's sound plays a role; a word with a difficult pronunciation is harder to learn than an easily pronounced one.

As for learning methods, there are some effective strategies, such as the *keyword* technique (Pressley et al, 1987) whereby an image of the word is suggested. Additionally, when learning a new word, vocalization – actually saying the word aloud – greatly helps memorization. Learning methods vary according to the student's level. Beginners learn according to the sound of the word, whereas more advanced students think of the meaning.

The relative effectiveness of vocabulary learning methods has been studied. In his research, Nation found that the "old fashioned" learning of words from lists, was not necessarily less effective than the "new" leaning in context method. It must be borne in mind, however, there are two ways of investigating learning methods. One is to conduct carefully designed "laboratory tests" in which, for example, some students are asked to learn a list of nonsense words in different ways, and their retention rates are compared. The problems with the laboratory methods were enunciated by Meara (1994). The fact that the "courses" are very short, the number of words used quite small, and the memorising techniques used very narrow in scope led him to be sceptical about the usefulness of this type of study.

As a result of these drawbacks another approach to learning method research seems necessary, that is to say the study of students' learning in the classroom. Although this seems less rigorous it is in fact closer to real-world language learning. Indeed many attempts have been made to observe real language students in order to find how they learn. This has produced lists of strategies and an attempt to classify them. For example Schmitt, in a series of articles has observed language learners in different real-life situations. In one case he collected data about students over a long period of years. (Schmitt 1998) He also elicited data from a group of Japanese students of English on their methods (Schmitt 1997), finding that while reference to dictionaries was the most popular approach, written or oral repetition were also used, and also that vocabulary learning methods changed between high school and adult students.

The acquisition of vocabulary in reading, especially in first languages has received considerable attention. In the 1980s Nagy and colleagues (for example Nagy, Anderson and Herman 1987) found that students increased their vocabulary unconsciously (incidentally) by reading. Similarly in second language learning, students' incidental learning has been investigated in various situations. One study (Day Omura and Hiramatsu 1991) found similar incidental learning as Nagy, and in a study of reading using dictionaries (Luppescu and

Day 1993) dictionary users retained more vocabulary than non-users.

If context helps readers to acquire vocabulary, perhaps strategies can be taught to develop students' conscious acquisition of vocabulary in context. This is known as inferencing meaning from context. Sternberg and Powell's (1983) theory of learning words from context distinguishes between external context, such as how often the unknown word occurs, and internal context, such as the prefixes and suffixes the words takes.

Inferencing from context also occurs when reading L2 texts. However, its effectiveness has been found to depend critically on the level of proficiency in the L2. If the *density* of unknown words (their frequency in relation to known words) is high then the reader will fail to infer the meaning of unknown words. Laufer's study of Israeli students led her to assert that students need a 3000 word-family vocabulary in order to be successful guessers. (Laufer 1997)

It is clear that while there are wide differences in L2 vocabulary learning, which depend on many factors, such as the relation between the L1 and L2, the students' general proficiency, age and possibly also gender, personality and motivation, intervention by an observant teacher can be effective. Improvements in the speed of learning and the length of retention and can be achieved by well-known strategies such as saying the word aloud, writing it down, using it in a written or spoken text and finding the L1 equivalent. However there is disagreement between teachers over whether effort put into strategy training is time well spent. As with all language learning, vocabulary learning is a demanding process, with many setbacks and false starts. As a result, even though some scepticism with regard to "miracle methods" may be warranted, one can sympathize with teachers who try to help students become effective learners.

Vocabulary research tests

There are two main purposes of vocabulary tests. One is to assess the size of a student's vocabulary and the other is to assess a student's overall proficiency. In the former type individual vocabulary items are elicited, using the discrete-point method. In the latter type sometimes known as integrative, vocabulary is tested in the context of longer reading or listening passages. Another terminology for the two types is decontextualized or context-independent tests, the latter contextualized or context-dependent.

Objective testing became very popular, especially in the USA. Spolsky (1995, 40) points out that the first tests of this kind appeared from 1916 onwards, gradually replacing earlier essay-type examinations. They reached their peak of popularity in the thirty years following Lado's book Language Testing in 1961. This was a time of great expansion in higher education, especially in the USA, and thus a time of great demand for efficient large-scale examinations for university entrance. The attraction of objective tests was based one type of test item: multiple choice. A recent example from the STEP Eikin test (level three) is as follows (Eiken 2010):

> The T-shirt I bought yesterday is a little too long. I want to (   ) it for a smaller one.
>
> 1 spend 2 grow 3 follow 4 exchange

This type of item came into use for several convincing reasons. It was fairly easy and quick to write, it could be based on a specific vocabulary list, it could be marked by machines, it has high reliability, and it was held to be a good indicator of language ability.

However, although objective tests have remained popular they have been subjected to scrutiny by scholars. The resulting criticisms were summarized by Wesche and Paribakht (1996). Among these were that candidates could score well by a process of elimination only, that a very small sample of the candidate's vocabulary is tested, and that the reason for a wrong answer could be lack of syntactical knowledge rather than vocabulary.

We noticed that there are many different aspects to defining a word, and this is reflected in vocabulary knowledge. One student may know just one meaning of a word, while another may know several idiomatic expressions using this word. This difference has been caught in the concepts of depth and breadth of vocabulary knowledge. Breadth is relatively superficial knowledge of many words, whereas depth is a fuller understanding, a knowledge "of all the distinctions that would be understood by an ordinary adult under normal circumstances." (Anderson and Freebody 1981, 93)

As a result of this distinction, researchers have constructed separate tests for measuring breadth or depth. However, not all words are equally familiar even to a native speaker. Children know many words with one meaning and later in life they acquire more meanings. Accordingly we need to group words and word meanings by difficulty. Especially for the purpose of testing vocabulary depth we need to establish a hierarchy of vocabulary difficulty. For native speakers Nagy and Anderson (1984) grouped word families in five levels expressing the degree of *relatedness* of the derived form to the base form. Thus the base form *moon* is simply related to *moonlight*, whereas higher up the scale the base word *content's* meaning is unrelated to that of the derived word *contentment*. For most L2 vocabulary size tests West's (1953) General Service List, strangely dating from half a century ago, is the normal starting point.

Having established what we need to test, namely a representative sample of West's list, what type of test is suitable? At least three basic item types are available: multiple choice, as in the example above, matching each word to a synonym or definition and choosing the correct L1 equivalent.

Two major vocabulary size tests are the Vocabulary Levels Test (Nation

1990) and the Eurocentres vocabulary size test. Nation's test claims to assess a candidate's vocabulary by placing him or her in one of five levels: 2000, 3000, 5000 word and university level bands. The definition of each level is based on Thorndike and Lorge's list of word frequencies (1944). The format is word to definition matching. For example:

1 apply 2 elect 3 jump 4 manufacture 5 melt 6 threat.

choose by voting ___ become like water ___ make ___

In validation studies this test has been found to be robust, but with one caveat. Because many of the difficult words in English are Latin-based, Romance language speakers are favored. However, in spite of this it has been praised by Meara (1996, 38) as "the nearest thing we have to a standard test of vocabulary." The popularity of this test is shown by its later development into The Vocabulary Size Test (Nation and Beglar 2009).

The purpose of the Eurocentres vocabulary size test is slightly different. It is used as a placement test for language school classes. It has the great virtue of speed, since it can be taken at a computer terminal in ten minutes, with the results given immediately.

The format, known as *checklist* is surprising at first glance Each word appears on the screen and the candidate answers the question "Do you know this word?" by clicking on a *yes* or *no* box. To avoid cheating, an ingenious method is applied. Among the words are many (a third of all the items, in fact) which are fake items or non-words, such as *cokram* or *obsolation*; if the candidate claims to know these, then his or her score is reduced. In validation studies (Meara 1996, 43) it has been found that candidates claim to know certain non-words depending on their native language. Moreover another problem is that low level candidates score lower than they should. However the practicality and speed of the test have made it popular, and there are hopes that the problems can be solved.

Turning now to vocabulary depth, also called *quality of vocabulary*, a word associates test was developed by Read. By depth he understood the number of different meanings of each word. As a result the words are not very difficult but they are used in more difficult ways than in the vocabulary size test. The item format was as follows:

Choose the related words from the list

Edit

Arithmetic film pole publishing revise risk surface text

In this case the correct answers are *film publishing revise* and *text*

In order to validate the test, Read conducted both statistical analysis and think-aloud studies with the participants. He found that because of guessing, vocabulary had not been successfully evaluated. A resulting later version of the test was less ambitious in scope, focusing "specifically on the meanings and collocational possibilities of adjectives." (Read 2000, 185)

We see that, based on established word-frequency lists, vocabulary size can be measured in an economical way. Moreover, although linguists might dispute that vocabulary size is directly related to general proficiency, it has proved to be a practical tool for student placement. On the other hand, vocabulary depth or range is more difficult to assess. Possibly with more effort an improved and practical test could be produced. Indeed, the concept of vocabulary size is easier to define than *language proficiency* or *communicative proficiency* which are the constructs underlying the major commercial tests.

The future of vocabulary testing

The vast increase in the use of computers for communication has changed many aspects of our life, from commerce to social life, and language testing has been no stranger to this trend. Particularly two aspects of testing have changed: test delivery and vocabulary scholarship.

Since 1990 there have been several large scale tests mediated by computer. The TOEFL examination, which is required by US universities for international students, has an *iBT* (internet-based) version in addition to its *PBT* (paper based test) version. Initially a computer-based test (CBT) version started in 1998 and the iBT was added in 2005. In this test reading, writing, listening and speaking are tested separately. However, there are no separate grammar or vocabulary sections. These skills are included in the other sections and subsumed under criteria such as "effective communication" or "accuracy of expression". Thus vocabulary is tested in a contextualized and integrative way. Similarly, the Cambridge examination organization offers a wide range of computer-based ESOL proficiency tests from the low-level Key English Test (KET) to the professional-level Business English Test (BULATS). Again vocabulary and grammar are not tested specifically but they are covered in the criteria for evaluation in the writing and speaking sections, known as "lexical resources" and "range of grammatical expression". A more recent test is the Pearson Test of English (PTE) of which the PTE Academic is similar in purpose to TOEFL and IELTS. It is relatively short, taken in a single three-hour test session, and consists of three timed sections, Speaking and Writing, Reading, and Listening. Again it is computer based, and in addition all reporting is done through the internet.

Thus the internet is being used to make testing less time-consuming and inconvenient for both the taker and institutions. Additionally test making organizations are slowly learning to make use of the computer.

The second way in which computers are shaping the future of vocabulary testing is in the production of vocabulary lists. The raw material for more accurate vocabulary lists is the computer-based corpus. This is a large collection of texts. For example, the British National Corpus is a 100 million word collection of samples of written and spoken language, and the TIME magazine archive contains 100 million words of text of American English from 1923 to the present, as found in TIME magazine. In addition to the corpus, special software allows the reader to compile lists according to certain criteria such as word frequency or collocations.

Recently, special corpora have been created in order to empirically make lists in certain specific fields. Two obvious candidates are business and academic English. An academic word list was based on a 3.5 million word corpus of academic texts at a New Zealand University (Coxhead 2000). Hyland and Tse (2007) noted that the AWL was a humanities-only list, so created a more wide-ranging academic corpus in Hong Kong. Finally, drawing on the fact that academic vocabulary contains not just words, but multi-word phrases, a group at the University of Michigan drew up an *academic formulas list* from both spoken and written sources (Simpson-Vlach and Ellis 2010). The core list contains 207 formulae which are common in both speech and writing. For example, the top three written fomulae are *on the other hand, due to the fact that* and *it should be noted* and the spoken counterparts are *be able to, blah blah blah*, (sic) *and this is the*. This corpus-derived knowledge guides writers of academic tests such as TOEFL and IELTS.

Conclusion

In summary, vocabulary knowledge is a vital part of fluency in a second language. Especially in the productive skills of speaking or writing a wide and growing vocabulary gives the student confidence and builds motivation. As a result evaluation of vocabulary plays an important pedagogical role, by making available to the student precise data on the growth of his or her language resources. We have noted that a working definition of a word is a necessary preliminary step in assessing vocabulary size, and that the word family is a useful concept. Furthermore measuring vocabulary has many applications. For example, vocabulary size is a useful indicator of reading ability.

As for the vocabulary testing method, there has been a trend in recent years, as part of the communicative language learning movement, to integrate vocabulary testing into tests of each of the four skills, or indeed to conflate two or more skills in one test. Thus we find the integrative items types in TOEFL where listening reading and speaking or listening reading and writing are combined. For the future, especially as the search for economical and speedy testing continues, one can envisage vocabulary tests being used as efficient measures of overall language proficiency. As a result, it is surely no exaggeration to say that vocabulary testing will play an important part in the future of language testing.

Note

The author acknowledges with gratitude the receipt of an AGU research grant which allowed him to visit Dr John Read at the University of Auckland New Zealand in March 2010.

References

Anderson, R. C. and Freebody, P. (1981) Vocabulary Knowledge. In J. T. Guthrie (ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark DE: International Reading Association.

Coxhead, A. (2000) A new academic word list. *TESOL Quarterly*, 34, 2: 213–238. [10.3]

Chapelle, C. (1998) Construct definition and validity inquiry in SLA research. In Bachman, L. F. and Cohen, A. D., editors, *Interfaces between second language acquisition and language testing research.* Cambridge: Cambridge University Press, 32–70.

Eiken (2010): http://www.eiken.or.jp/listening/grade_3.html

Hylan, K. and Tse, P. (2007) Is there an "academic" vocabulary? *TESOL Quarterly*, 21, 235–253.

Laufer, B. (1997) *The lexical plight in second language reading, Second language vocabulary acquisition: A rationale for pedagogy*, Cambridge: Cambridge University Press, 20–34.

Laufer, B. (1997) What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In Schmitt, N. and McCarthy, M. (eds.) (1997) 140–155.

Luppescu, S. and Day, R. R. (1993) Reading, dictionaries, and vocabulary learning, *Language Learning*, 43, 263–287.

McKeown, M. G. and M. E. Curtis (eds.). (1987) *The Nature of Vocabulary Acquisition.* Hillsdale, NJ: Lawrence Erlbaum.

Meara, P. (1994) Second Language acquisition: lexis. In Asher, R. E. (ed.). *The Encyclopedia of language and linguistics*, Vol. 7 (pp. 3726–3728), Oxford: Pergamon Press.

Meara, P. (1996) "The dimensions of lexical competence". In G. Brown, K. Malmkjaer and J. Williams (eds.), *Performance and Competence in Second Language Acquisition.* CUP. pp. 35–53.

Nagy, W. E., Anderson, R. C. and Herman, P. (1987) Learning word meanings from context during normal reading. *American Educational Research Journal*, 24, 237–270.

Nation, I. S. P. (1982) Beginning to learn foreign vocabulary: A review of research. *RELC Journal*, 13, 14–36.

Nation, P. (1990) *Teaching and learning vocabulary.* New York: Newbury House.

Nation, P. and Beglar, D. (2007) A vocabulary size test. *The Language Teacher*, 31, 7, 9–12.

Nattinger, J. and DeCarrico, J. (1992) *Lexical phrases and Language Teaching.* Oxford University Press.

Pressley, M., J. R. Levin and M. A. McDaniel (1987) Remembering vs. inferring what a word means: Mnemonic and contextual approaches. In McKeown and

Curtis (eds.), (1987) pp. 107–127.

Quinion, M. “How many words?” http://www.worldwidewords.org/articles/howmany.htm

Read, J. (1993) The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10: 3, 355–371.

Read, J. (2000) *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Read, J. and Chapelle, C. A. (2001) A framework for second language vocabulary assessment. *Language Testing*, 2001, 1.

Richards, J. (1976) The role of vocabulary teaching. *TESOL Quarterly*, 10, 77–89.

Schmitt, N. (1997) Vocabulary learning strategies. In Schmitt and McCarthy 1997: 199–227.

Schmitt, N. (1998) Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, 48 (2), 281–317.

Schmitt, N. and McCarthy, M. (eds.) (1997) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press.

Simpson-Vlach, R. and Ellis, N. C. (2010) An Academic Formulas List (AFL). *Applied Linguistics*, 31, 487–512.

Sternberg, R. J. and Powell, J. S. (1983) Comprehending verbal comprehension. *American Psychologist*, 38: 878–893.

Wesche, M. and Paribakht, T. S. (1996) Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53, 13–40.