



Second Language Vocabulary Assessment: Current Practices and New Directions

JOHN READ*
University of Auckland

ABSTRACT

This paper surveys some current developments in second language vocabulary assessment, with particular attention to the ways in which computer corpora can provide better quality information about the frequency of words and how they are used in specific contexts. The relative merits of different word lists are discussed, including the Academic Word List and frequency lists derived from the British National Corpus. Word frequency data is needed for measures of vocabulary size, such as the Yes/No format, which is being developed and used for a variety of purposes. The paper also reviews work on testing depth of knowledge of vocabulary, where rather less progress has been made, both in defining depth as a construct and in developing tests for practical use. Another important perspective is the use of vocabulary within particular contexts of use or registers, and recent corpus research is extending our understanding of the lexical features of academic registers. This provides a basis for assessing learners' ability to deploy their vocabulary knowledge effectively for functional communication in specific academic contexts. It is concluded that, while current tests of vocabulary knowledge are valuable for certain purposes, they need to be complemented by more contextualised measures of vocabulary use.

KEYWORDS: vocabulary tests, assessment, corpus analysis, word frequency, registers, vocabulary size, depth of vocabulary knowledge, contexts of use.

* *Address for correspondence:* Associate Professor John Read. Department of Applied Language Studies and Linguistics, University of Auckland, Private Bag 92019, Auckland 1132, New Zealand. Tel: +64-9-373 7599 Extn 87673; Fax: +64-9-308-2360. E-mail: ja.read@auckland.ac.nz.

I. INTRODUCTION

In some respects vocabulary testing is quite a simple activity, a matter of selecting a suitable number of target words and assessing whether each one is known by means of an established test format such as multiple-choice, matching, gap-filling, or some form of translation. Such tests continue to be routinely used in second language teaching for a variety of assessment purposes and, if well designed, can be highly reliable and efficient measures of learner competence. However the dominance of the communicative approach to language teaching in the past thirty years has thrown up various challenges to the validity of the conventional vocabulary test and this has prompted some re-thinking of the nature of lexical ability as well as how it can best be assessed. The most comprehensive discussion of the issues can be found in Read's (2000) book on vocabulary assessment, and this article will give particular attention to further developments since that book was written.

One main theme running through the article is the actual and potential contribution made by corpus analysis to vocabulary assessment. The development of computer corpora has had an enormous influence on vocabulary studies of all kinds, most notably in the routine use of corpus evidence by lexicographers in defining words, recording their frequency of occurrence, illustrating typical uses of the words and so on. Corpus analysis has also offered new insights into the collocational behaviour of words and the functioning of multiword lexical items in written and spoken discourse. For vocabulary assessment corpora can provide the basis for more accurate word lists from which target words can be sampled, taking account of frequency, range of occurrence and other criteria. In addition, corpus analysis yields descriptions of the lexical features of language as it is employed in specific contexts of use, such as academic disciplines. The applications of such descriptions for the assessment of learners' lexical abilities have yet to be fully explored.

The article first considers work on measuring breadth of vocabulary knowledge among second language learners, through estimating the number of words that they know. Since this usually involves the use of word frequency lists to provide a sampling frame, the suitability of various accessible lists is considered. This leads to a discussion of one particularly interesting method of estimating vocabulary size, the Yes/No format. The second topic in the article is testing depth of vocabulary knowledge, using methods such as the word associates format and the Vocabulary Knowledge Scale to assess how well particular words are known by learners. The third main section takes up the issue of how to assess vocabulary knowledge in context. It draws on recent corpus research, which is providing much better descriptions than were previously available of the lexical features of particular registers, which are varieties of

language belonging to specific contexts of use (in this case, academic disciplines). The article concludes with suggestions as to how these corpus-based descriptions can inform the assessment of vocabulary knowledge from a sociolinguistic perspective.

II. MEASURING VOCABULARY SIZE

One longstanding area of research and development in second language vocabulary assessment is the estimation of vocabulary size (often referred to as breadth of vocabulary knowledge). There are several purposes for making such estimates, both for native speakers and for learners. For instance, since vocabulary size is closely associated with reading comprehension ability, vocabulary tests have traditionally had a significant role in research on reading development and in literacy programmes. For second language learners, vocabulary assessment can reveal the extent of the lexical gap they face in coping with authentic reading materials and undertaking other communicative tasks in the target language.

Vocabulary size measures typically require a relatively large sample of words that represent a defined frequency range, together with a simple response task to indicate whether each word is known or not. Let us look at each of these aspects in turn.

II.1 Sampling from Word Frequency Lists

There is an extensive literature on the vocabulary size of native speakers of English, which has produced widely varying estimates of the number of words that they know. Much of the earlier research was characterised by methodological weaknesses (see, eg, Lorge & Chall, 1963; Nation, 1993), such as confusion over what constituted a word and sampling procedures which led to high frequency words being overrepresented. Some later studies have yielded more realistic estimates based on careful definition of what a “word” is and well-designed sampling procedures (Goulden, Nation, & Read, 1990; Zechmeister *et al.*, 1995). Normally for native speakers the sample of words to be tested has come from a large dictionary of contemporary English in order to cover as many as possible of the words that the participants in the study are likely to know.

One limitation of dictionaries for native-speaker users is that they do not give any explicit information about the frequency of words, because it is generally preferable to present the words in a vocabulary size test in ascending order from most common to rare. This deficiency in dictionaries is overcome by word frequency lists that are based on computer corpora. The most accessible lists of this kind are those compiled from the British National

Corpus by Leech, Rayson and Wilson (2001), which are available both in book form and from a companion website. There are separate lists of word forms and of lemmas, as well as lists for both the written and spoken sub-corpora of the BNC.

However, the preferred lexical unit for current vocabulary size research is the word family, which consists of a base word and its inflected forms, together with derived forms which share a common meaning with the base word. Two examples are *differ*, *differs*, *differed*, *differing*, *different*, *differently*, *difference*, *differences*; and *rich*, *richer*, *richest*, *richly*, *riches*, *richness*. Thus, Nation has reworked the Leech *et al.* (2001) data into about 14 1000-item lists of word families, following the largely morphological criteria for identifying word family members developed by Bauer and Nation (1993). Nation's lists can be downloaded from his website, <http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx>, where one version of a Vocabulary Size Test based on the lists can also be accessed. Another version is published in Nation and Gu (2007).

Vocabulary size tests for second language learners understandably focus on a narrower range of words than those for native speakers, since low frequency words are much less likely to be known, especially by learners in a foreign language environment. The aging General Service List (GSL) (West, 1953) still provides a foundation for work in this area, with its selection of 2000 high-frequency word families, which account for a high percentage of the running words in any written or spoken English text, now as they did fifty years ago. The list is often criticised for its outmoded entries and its lack of modern terms, especially among the second 1000 words, but it has yet to be replaced by any more up-to-date compilation that would draw on the best of contemporary corpus data while retaining the GSL's sound selection criteria of frequency, range, familiarity and pedagogical value (Nation & Waring, 1997; Read, 2000: 227-28).

One list that does combine these virtues is the Academic Word List (AWL), Coxhead's (2000) set of 570 word families occurring frequently in written texts across a range of university disciplines. The AWL has been very influential in recent years in the teaching and testing of English for academic purposes as a reference list for the sub-technical vocabulary that students are assumed to need in undertaking university studies through the medium of English. The AWL complements the GSL in the sense that words appearing in the general list were excluded from the academic list. In EFL countries such as Indonesia (Nurweni & Read, 1999) and Japan (Beglar & Hunt, 1999), researchers have conducted studies which presuppose that tests based on samples of words from the GSL and the AWL

will reveal most of what can usefully be elicited about the English vocabulary size of first-year university students in those countries.

The AWL itself is not above criticism. It is based on the assumption that students participating in an English for Academic Purposes programme are intending to study in a range of disciplines and therefore the vocabulary they learn should represent a common core of words. Ward (1999) argues that a different approach is required if all the students are entering a particular department or faculty. Taking the case of engineering students at a Thai university, Ward compiled a list of the 3000 most frequent word families in their first-year textbooks (written in English) and found that the first 2000 of them provided substantially better coverage of the lexical content of engineering textbooks than did the General Service List plus the University Word List (a precursor of the AWL). Since reading their textbooks was these students' primary need to use English, Ward's argument was that it was more cost-effective for them to study the specialised Engineering Word List rather than working with a general academic list like the AWL. This assumes of course that specialised lists are available for various academic disciplines, which is generally not the case.

A more comprehensive critique of the AWL has just been published by Hyland and Tse (2007). First, they provide evidence that Coxhead's (2000) AWL corpus was biased in favour of business studies and law, while underrepresenting the natural sciences and engineering. Using their own corpus of academic texts, Hyland and Tse found that most AWL word families were not very frequent overall and occurred very unevenly across the three disciplinary areas in their corpus: Sciences, Engineering and Social Sciences. These authors go on to argue that, even where word families are found in a range of fields, the meanings of the words and the ways they collocate are quite distinctive in each discipline. This leads Hyland and Tse to question the value of any vocabulary list that attempts to specify a common core of academic words, especially if it takes no account of meanings and collocational preferences. We will return to this issue of the discipline-specific uses of vocabulary later.

The basic point to be taken from this discussion of various lists is that there is no definitive word frequency list, either for English generally or for particular uses of English. In drawing a sample of words for a vocabulary size test, it is very much a case of using the best available (or perhaps, least unsatisfactory) list. Even though computers make it easy to generate data on the frequency of word forms in a particular corpus, a great deal more work is needed to develop a well-formulated word list as the basis for a vocabulary size test, especially if word meanings need to be taken into account. An alternative approach, proposed

particularly for languages other than English (where typically little if any word frequency data is available), is to rely on the judgement of language teachers or other linguistic experts. For example, Daller, van Hout and Treffers-Daller (2003: 208-210) obtained reliable frequency judgements for Turkish words in this manner. However, Alderson (2007) found that linguists at a British university achieved only moderate success in ranking sets of English words according to their frequency in the British National Corpus. Further investigation is needed to determine whether, and in what situations, reliable estimates of frequency can be obtained.

II.2 The Yes/No Format

Once a suitable list has been chosen, the next step in developing a vocabulary size test is to select a sample of target words for the test items. Since a relatively large sample is required to make reliable estimate of vocabulary size (Nation, 1993), test designers tend to prefer a simple test format. The most widely used measure of English vocabulary size for second language learners, Nation's Vocabulary Levels Test (Nation, 2001: 416-424; Schmitt, Schmitt & Clapham, 2001), requires the test-takers to match words with their synonyms or short definitions. Nation's new Vocabulary Size Test (Nation & Gu, 2007), referred to above, has a multiple-choice format, with each target word presented in a short non-defining sentence followed by four possible definitions as options. These two types of item provide direct evidence that each word is actually known.

The simplest test format was originally called the checklist and is now generally referred to as the Yes/No format. In this case the test-takers are presented with a series of words and just indicate whether they know each word or not. Anderson and Freebody (1983) introduced an important innovation by including among the items a substantial proportion of non-words. This provided a basis for adjusting the scores of test-takers who responded "Yes" to a significant number of non-words, on the assumption that such learners were overestimating their vocabulary knowledge. There are interesting technical issues in devising a scoring system for the test that achieves a satisfactory balance between giving adequate credit for self-reported knowledge of real words and imposing an appropriate penalty for claimed knowledge of non-words. Several studies (Beeckmans *et al.*, 2001; Huibregtse, Admiraal & Meara, 2002; Mochida & Harrington, 2006) have investigated a number of mathematically complex scoring formulas. However, it appears that in practical testing situations, if it can be assumed that the test-takers are honestly reporting most of the time whether they know the target words or not, a simple calculation such as the number of Yes

responses to real words minus the number of Yes responses to non-words yields a reasonably valid measure of vocabulary size.

Meara and his colleagues, first in London (eg Meara & Buxton, 1987) and then at Swansea University (Meara, 1992; Meara & Milton, 2005) have taken the lead in developing Yes/No tests for second language learners and making them available for practical use as placement tests or as general measures of vocabulary size or competence in the language. Since the Yes/No format lends itself well to computer administration, the current Swansea tests are available commercially on CD-ROM (Meara & Milton, 2005) or as free downloads from the Swansea website (Meara & Miralpeix, 2006). One program, X_Lex, covers the first 5000 most frequent words of not only English but also French, Spanish, Swedish and Portuguese, whereas Y_Lex samples vocabulary in the 6000—10,000 word range, but just for English at present (see Miralpeix, this volume).

A further development is a version of X-Lex in which the words are presented not in written form on the screen but orally. This test, known as Aural_Lex (Milton & Hopkins, 2005), is notable in the first instance because there are so few tests of spoken vocabulary available, reflecting a more general neglect of words in their spoken form in vocabulary studies (cf. Read, 2000: 235-239). Ideally, tests like Aural_Lex should be based on word frequency data from spoken language corpora. On the other hand, Milton and his co-researchers (Milton & Hopkins, 2006; Milton & Riordan, 2007) have exploited the parallel lexical content of X_Lex and Aural_Lex in two studies to test their hypothesis that Arabic speakers learning English might underestimate their vocabulary knowledge in a written test through faulty recognition of the target words resulting from the fact that they transfer their consonant-based decoding strategies from Arabic to English. However, both studies showed that Arabic-speaking learners were not disadvantaged by being presented with words in their written rather than spoken form.

More investigation is needed of how recognisable isolated words are when they are presented orally to listeners, regardless of their L1 background. As Milton and Riordan (2007: 132) note, whereas the printed form of a word is relatively fixed, the spoken form can vary according to factors such as the linguistic context, the speaker's accent and the potential for a word in isolation to be confused with a similar-sounding word. With this in mind, the present author is currently involved in a project to develop Yes/No tests in which spoken words are associated with two kinds of sentence context: one providing a lexically bare syntactic context and the other a semantically richer one (Read, 2007). The contexts may add to the basic format not only a more accurate identification of the target word but also a link to a specific

use of the word, which could result in more valid judgements by the test-takers as to whether they know the word or not.

Another significant application of the Yes/No format is found in DIALANG (www.dialang.org), the web-based system through which learners of 14 European languages can assess their proficiency in the target language, in terms of the Common European Framework of Reference (CEFR) (Council of Europe, 2001). When learners access the system, they are invited to assess their own skills in the target language and also to take a Vocabulary Size Placement Test (VSPT) in the Yes/No format. The purpose of these two measures is to determine the general proficiency level of the learners so that, when they take a specific skill test, the system will present them with items and texts that are broadly suited to their level. According to Alderson's (2005: 79-96) analysis of pilot test data, the VSPT for English performs this role effectively, as shown in these very substantial correlations with the English skills tests in DIALANG:

Reading	.64
Grammar	.64
Writing	.70
Listening	.61
Vocabulary skills	.72

Apart from presenting an account of the development and validation of DIALANG, Alderson (2005) explores in a more general manner the concept of diagnosis in language assessment. Diagnosis as a purpose for language tests has been relatively neglected by researchers and test developers, in favour of tests of proficiency and communicative performance. A new interest in diagnosis in language testing is likely to prompt renewed attention to the role of vocabulary—as well as grammar—tests as measures of learners' knowledge of the structural system of the target language, to complement skills-based assessments of their L2 proficiency.

Thus, there are well-established procedures for measuring the size of a learner's vocabulary. It is difficult to find the "perfect" word frequency list to provide a suitable sample of lexical items, although various options are now available for selecting words. Since quite a large sample of words is required to make a reliable estimate of vocabulary size, simple types of test item are preferred. Despite its simplicity, the Yes/No format has proved to be an

informative and cost-effective means of assessing the state of learners' vocabulary knowledge, particularly for placement and diagnostic purposes.

III. TESTING DEPTH OF KNOWLEDGE

Compared to these robust initiatives to develop and apply various tests of vocabulary size, there has been rather less progress in measuring quality (or depth) of vocabulary knowledge. The case for testing depth is built on the recognition that, whereas a size test typically assesses the learner's ability to associate the written form of a word with a simple statement of its meaning, there is in fact much more to know about words if they are to become functional units in the learner's L2 lexicon: how the word is pronounced and spelled, what its morphological forms are, how it functions syntactically, how frequent it is, how it is used appropriately from a sociolinguistic perspective, and so on. Numerous authors (eg Henriksen, 1999; Nation, 2001; Read, 2004) have analysed the components of word knowledge and discussed how they can be assessed. It is generally acknowledged that, except for certain research purposes (see, eg, Schmitt, 1998), there is little point in eliciting all that learners may know about a particular set of words. This means that there is a role for measures which focus selectively on key aspects of word knowledge.

One type of test that has been adopted to some extent is Read's (1993, 1998) word associates format. This builds on the concept of word association by creating items that consist of a target word and six or eight other words, half of which are associated with the target word and half not. The relationships between the words are primarily semantic and collocational, and the format offers opportunities to assess some key elements of the core meaning of the target word, or alternatively more than one meaning of the word. The aim was to design a simple type of item that would test deep word knowledge in a meaningful way.

Read's initial studies involved learners studying English for academic purposes at a New Zealand university. Several scholars in the Netherlands have developed modified versions of the format and introduced innovations in the design of the items. Some (eg, Bogaards, 2000; Greidanus & Nienhuis, 2001; Greidanus, Beks & Wakely, 2005) have found the test to be a challenging measure of vocabulary knowledge for advanced foreign language learners at university level. On the other hand, Schoonen and Verhallen (in press) employed the test in primary schools in the Netherlands to explore the extent to which the lexical development of children for whom Dutch was a second language lagged behind that of their native-speaking peers. In Canada, Qian (1999, 2002) used Read's (1998) test in his studies of

the relationship between L2 vocabulary knowledge and reading comprehension ability among adult learners of English. More recently, Qian and Schedl (2004) concluded that word associates items would be a feasible alternative to conventional multiple-choice items as measures of vocabulary knowledge in the Test of English as a Foreign Language (TOEFL).

Another measure of deep word knowledge which has gained some currency is Paribakht and Wesche's (1997) Vocabulary Knowledge Scale (VKS). These researchers were interested in the "incidental" acquisition of word meaning through intensive reading activities. Needing a means of recording partial understanding of their target words, they devised a six-point elicitation scale which went from "I don't remember having seen this word before" to "I can use this word in a sentence". Thus, the scale combines self-report with some verifiable evidence of word knowledge in the form of a synonym, L1 translation or sentence. The learners use the scale to report how well they know each of the target words. Other researchers who have used modified versions of the VKS in their studies of L2 vocabulary development include Joe (1998) and Zareva, Schwanenflugel and Nikolova (2005).

Some other tests of depth of knowledge could be mentioned here but there is certainly no depth measure that has reached the same wide acceptance that the Vocabulary Levels Test has achieved as a measure of vocabulary size. As Read (2004) has pointed out, the depth of knowledge construct is more diffuse than the vocabulary size (or "breadth") construct. There are so many more aspects of word knowledge that could potentially be assessed and no consensus has emerged as to which are the most significant ones. In fact, some authors (eg Vermeer, 2001) have argued that learners' word knowledge naturally deepens as vocabulary size increases, so that good size measures may be all that are required. It remains to be seen whether particular tests will be adopted as standard measures for practical use in language teaching, rather than as research instruments, which most tests of this kind are at present.

IV. ASSESSING VOCABULARY USE IN CONTEXT

The vocabulary tests discussed so far have all presented the target words as isolated lexical units with no reference to context. The issue of whether words should be assessed in context is a longstanding one in vocabulary studies (Read, 2000: 99-115, 161-165), and a range of commonsense arguments can be put forward in favour of one position or the other. Certainly, the dominant communicative approach to language teaching and testing calls into question the notion that decontextualised learning of language forms is the basis for effective proficiency development in a second language. Hyland and Tse's (2007) critique of the notion of a

general academic vocabulary, as cited above, very much reflects this view that learners should engage with the actual use of lexical items in specific contexts if they are to be successful language users in the academic environment or elsewhere.

From the perspective of current theory in the area of test validation, Read and Chapelle (2001) put forward a somewhat similar argument in their framework for second language vocabulary assessment. They noted that most existing vocabulary tests implicitly defined vocabulary knowledge as a *trait*, a mental attribute of the learner that could be described and measured without any reference to the contexts in which the words are used. Obviously, vocabulary size measures such as the Yes/No format discussed above represent a classic example of this approach to assessment. Read and Chapelle argued that such measures need to be complemented by others—based on an *interactionalist* definition of vocabulary—which would assess the learners' ability to deploy their vocabulary knowledge appropriately in particular contexts of use. In the article they gave only a general indication of what kind of measures could be used for this purpose and how they might be developed. One specific example was vocabulary size in mathematics, which shows that they were thinking in terms of variation in lexical use according to academic discipline.

If we pursue this notion of discipline-specific vocabulary use, we need to have ways of, first, defining the appropriate divisions of academic discourse and then identifying the distinctive lexical features of each division. In the first case, there are numerous divisions that can be made at varying levels of generality: broad distinctions between the humanities, law, business studies, physical sciences, and engineering; down to disciplines such as psychology, accounting, chemistry, physiology, and English literature; and on to a whole range of sub-disciplinary fields. A thriving literature in the field of EAP explores differences in academic discourse across disciplines (eg, Hyland, 2000; Swales, 1990). In addition, an increasing number of computer corpora are available to document the features of academic discourse, both general corpora that include a sub-corpus of scholarly texts and smaller specialised corpora which comprise academic texts of various kinds.

In discussing particular disciplines or contexts of use, we need an appropriate way to designate a language variety associated with each one. The term *register* is useful for this purpose because it has quite a long history, particularly in British applied linguistics. It was given fresh currency in the influential corpus-based research of Biber and his associates (Biber, 1988; Biber & Finegan, 1994). The term is a suitable one to use when a language variety is defined in terms of its lexical and grammatical features, although the existence of such varieties is a matter of some debate (see, eg, Davies, 2001; Douglas, 2005).

One way to approach the distinctive features of academic registers is through the study of technical vocabulary. This is an area that has not received much scholarly attention until recently. Obviously there are many dictionaries on the market for a whole range of disciplines, compiled by authors with suitable subject-area expertise for the benefit of students and other novices in each field. However, we have lacked systematic procedures for identifying the technical words in particular texts.

This gap is now being addressed in two ways. One, exemplified in the work of Chung and Nation (2003), involves the use of judgements based on expertise in the appropriate subject area. Chung and Nation developed a rating scale to classify all the words in a university textbook into four categories according to their degree of technicalness. The two textbooks analysed for the study were in anatomy and applied linguistics, both fields in which Chung has university qualifications. Figure 1 gives examples of words in each of the four categories from the anatomy textbook. The results of the lexical analyses showed that technical words accounted for 31% of the running words in the anatomy text (at Steps 3 and 4 of the rating scale), as compared to a figure of 21% for the applied linguistics book. These figures were much larger than had previously been suggested (eg, by Nation, 2001: 12). Obviously, this kind of painstaking manual analysis is too time-consuming to be practical for widespread use.

<p>Step 1 Words with no semantic relationship to anatomy: <i>the, is, between, amounts, common, directly</i></p>
<p>Step 2 Words whose meaning is minimally related: <i>superior, part, forms, pairs, structures, surrounds</i></p>
<p>Step 3 Words whose meaning is closely related to anatomy but also in general use: <i>chest, trunk, neck, abdomen, ribs, breast</i></p>
<p>Step 4 Words with a specific meaning in anatomy, not used in general language: <i>thorax, sternum, costal, pectoral, fascia, periosteum, viscera</i></p>

Figure 1. A rating scale for finding technical words (as applied to an anatomy text)
(Chung & Nation, 2003: 105)

The alternative approach, then, is to apply the tools of corpus analysis to the identification of vocabulary that is characteristic of particular texts or registers. One early proposal along these lines was put forward by Yang (1986), who located both single-word and multi-word technical terms in science texts by searching for items with “peak” frequency

in one field and little if any occurrence elsewhere. The same principle is the foundation for the Keyword analysis in *WordSmith Tools* (Scott, 2004), a widely used package for corpus analysis. The basic statistic is a chi-square analysis of the difference in the frequency with which a word occurs in a specific text or subcorpus as compared to a general reference corpus. Thus, on the basis of this statistical criterion, the key words are assumed to be those that occur much more frequently in a specified type of text than in the language generally.

Chujo and Utiyama (2006) have pushed the keyword concept a step further in their research on the vocabulary of business English. Taking the commerce and finance section of the British National Corpus as their sample of business English, they applied no fewer than nine statistical measures to this sub-corpus. It is beyond the scope of this article to consider all nine statistics here (and in fact some pairs of measures produced very similar results). Table 1 shows the 20 most frequent words identified by three of the statistics: the complementary similarity measure (CSM), chi-square (χ^2) and mutual information (MI). One limitation of the CSM list is that over half of the most frequent items are function words, although it also includes nouns such as *company*, *market* and *business*. By contrast, the chi-square list is, with one exception, composed of content words which are recognisably associated with business. These words look as if they belong in the third category of Chung and Nation's (2003) scale, in that they are closely related to business topics but also in general use. By contrast, the MI statistic identified not only all content words but also items that are more distinctly technical. Terms such as *lading*, *arbitrage*, *offeror* and *settlor* would appear to belong in the fourth category on the Chung and Nation scale, since they are not in general use outside a business context.

Thus, Chujo and Utiyama's (2006) work is promising, in that it may provide an automated alternative to Chung and Nation's (2003) rational basis for identifying degrees of technicalness in the vocabulary of a particular register. Further research is desirable in other subject areas but it appears that a judicious application of corpus statistics—with suitable editing—could yield a practical method of both developing a reference list of technical terms for a register and identifying the distinctive lexis of specific texts. This would then provide a basis for designing appropriate assessment procedures.

Complementary Similarity	Chi-square	Mutual Information
-------------------------------------	-------------------	---------------------------

the	market	lading
of	company	buyout
be	bank	long-run
to	price	arbitrage
a	business	subcontractor
in	investment	stockmarket
will	rate	offeror
for	firm	drafter
company	cost	no-arbitrage
market	rate	shareholding
by	account	headhunter
or	the	payout
business	profit	issuer
this	contract	liquidity
may	share	salesperson
bank	income	settlor
price	customer	acquirer
cost	asset	volatility
which	financial	accountancy
rate	investor	lender

Table 1. 20 most distinctive words in the BNC Commerce and Finance sub-corpus, as calculated by three measures (from Chujo & Uityama, 2006: 262)

Another perspective on the registers of academic English is found in the work of Biber (2006) and his colleagues on the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) Corpus. The corpus was commissioned by the Educational Testing Service as part of the developmental research on what has become the internet-based Test of English as a Foreign Language (iBT) (www.ets.org/toefl), in order to identify patterns of language use in university registers, and to help validate input texts for the iBT listening and reading tasks. There are some interesting findings concerning vocabulary use in the T2K-SWAL Corpus. Biber (2006) made a broad comparison between the spoken language of university classroom teaching and the written language of the textbook. Not surprisingly, perhaps, classroom teaching was characterized by the extremely high frequency of very common words like *get*, *say*, *think*, *want*, *see* and *thing*, whereas the textbooks contained a wide range of word types that occurred with low frequency, especially nouns with specialized meanings such as *disillusionment*, *enhancement*, *globalization*, *hominid* and *locus*.

An analysis across disciplinary areas showed some variation in the general pattern, as presented in Figure 2. Business and Engineering used relatively fewer word types than the

other disciplinary clusters in classroom teaching as well as in the textbooks. Presumably this is partly because these two subject areas are quite integrated and composed of allied fields of study, as compared to the more diverse range of disciplines included in the other three clusters. Nevertheless, it is also noticeable in Figure 1 that the disparity in word types between spoken and written academic language gets progressively larger as we move to the right of the graph, to the extent that in the Social Sciences and Humanities there are twice as many word types used in the textbooks as in classroom teaching.

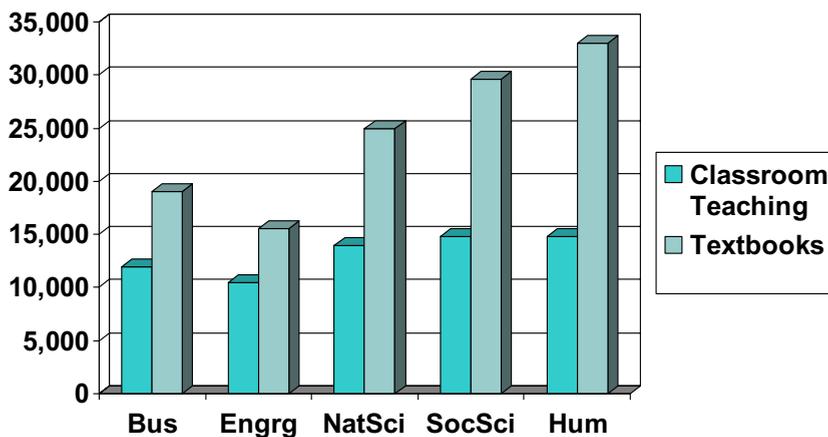


Figure 2. Word types in the T2K-SWAL Corpus by discipline (from Biber, 2006: 42)

Thus, research in corpus linguistics in particular is providing new analytical tools and insights that allow us to describe the lexical features of academic and other registers in more detail than was previously possible. Assuming that a relevant corpus is available or can be compiled (which are both significant provisos at this point), there are several ways in which such descriptions might contribute to discipline-specific vocabulary assessment.

- The most obvious contribution is in generating a list of lexical items that could be used to assess vocabulary competence in relation to a particular register. The raw lists from the output of corpus analysis are likely to need editing according to various criteria, including decisions about which categories of words (in terms of Chung and Nation's (2003) scale of technicalness) should be included, and a checking procedure to ensure that frequent words also occur across a range of texts within the register.

- Both vocabulary size and depth of lexical knowledge could be tested within particular registers or disciplines.
- However, taking up a point made by Hyland and Tse (2007), it would be highly desirable to assess the ability of learners to interpret the semantic and pragmatic features of the target lexical items as they are used in specific-purpose texts.
- Learners also need to be assessed on their ability to use specific-purpose vocabulary appropriately in academic speaking and writing tasks, including knowledge of typical collocations of words in disciplinary discourse.

The latter two suggestions take us beyond vocabulary tests of the conventional variety. To use the terminology adopted by Read (2000), we need to complement discrete, selective and context-independent tests such as the Yes/No and word associates formats with embedded, comprehensive and context-dependent measures that assess vocabulary knowledge and use through performance tasks of various kinds. In listening and reading tasks that would mean including questions which measure contextual understanding of lexical items in the text, whereas in speaking and writing assessment it involves applying range and appropriateness of vocabulary use as one of the criteria for rating the learners' performance. These ideas are not new but the point is that register-specific lexical analyses can allow test developers and raters to make better informed decisions about which aspects of the vocabulary they should focus on, and thus lead to more valid assessments.

V. CONCLUSION

This review of current developments illustrates how the twin strands of vocabulary assessment identified by Read (2000), represented here by discrete Yes/No tests and embedded measures of lexical production in specific academic contexts, are both still influential in their own ways. On the one hand, the construct of vocabulary size is a valuable means of assessing not only the state of learners' lexical knowledge but also their linguistic competence in a broader sense, for such purposes as placement in a language teaching programme and diagnosis of learning needs. The validation results from DIALANG even suggest that vocabulary size is associated with communicative proficiency to a considerable degree. The related construct of depth of vocabulary knowledge is less well defined and its practical applications for assessment purposes are less certain, but there is evidence that a well-designed test such as the word associates format has a useful role in at least some

educational contexts as a means of probing the semantic richness of the learner's mental lexicon.

On the other hand, in the field of language teaching, where it has become commonplace to define both goals and methods primarily in communicative terms, it can seem rather suspect to be assessing vocabulary knowledge through decontextualised test items. We need to complement discrete vocabulary tests with embedded measures of the learners' ability to handle lexical items in context. Traditionally, context has been conceived in linguistic terms as a sentence or larger co-text in which a vocabulary item occurs. Corpus analysis now gives us powerful tools to enrich our understanding of context through providing detailed descriptions of vocabulary use in specific registers or disciplinary genres. The applications of these richer descriptions to vocabulary assessment are not entirely clear as yet but they represent fertile ground for further investigation.

REFERENCES

- Alderson, J.C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.
- Alderson, J.C. (2007). Judging the frequency of English words. *Applied Linguistics*, 28, 383-409.
- Anderson, R.C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Huston (Ed.), *Advances in reading/language research*. Volume 2. Greenwich, CT: JAI Press, pp. 77-117.
- Bauer, L., & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6, 253-279.
- Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M. & Van de Velde, H. (2001). Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, 18, 234-274.
- Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and the university word level vocabulary tests. *Language Testing*, 16, 131-162.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., & Finegan, E. (Eds.) (1994). *Sociolinguistic perspectives on register*. New York: Oxford University Press.

- Bogaards, P. (2000). Testing L2 vocabulary at a high level: The case of the Euralex French tests. *Applied Linguistics*, 21, 490-516.
- Chujo, K., & Utiyama, M. (2006). Selecting level-specific vocabulary using statistical measures. *System*, 34, 255-269.
- Chung, T.M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, 15, 103-116.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34, 213-238.
- Council of Europe (2001). *A common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24, 197-222.
- Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing*, 18, 133-147.
- Douglas, D. (2005). Testing languages for specific purposes. In E. Hinkel (Ed.), *Handbook of research in second language teaching*. Mahwah, NJ: Lawrence Erlbaum, pp. 857-868.
- Goulden, R., Nation, P. & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 10, 341-363.
- Greidanus, T., Beks, B., & Wakely, R. (2005). Testing the development of French word knowledge by advanced Dutch and English speaking learners and native speakers. *Modern Language Journal*, 89, 221-233.
- Greidanus, T. & Nienhuis, L. (2001) Testing the quality of word knowledge in L2 by means of word associations: Types of distractors and types of associations. *Modern Language Journal*, 85, 567-577.
- Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, 21, 303-317.
- Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no test: Correction for guessing and response style. *Language Testing*, 19, 227-245.
- Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow, UK: Longman.
- Hyland, K. & Tse, P. (2007). Is there an “academic vocabulary”? *TESOL Quarterly*, 41, 235-253.
- Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics*, 19, 357-377.

- Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman. See also the companion website at: <http://www.comp.lancs.ac.uk/ucrel/bncfreq/>
- Lorge, I., & Chall, J. (1963). Estimating the size of vocabularies of children and adults: A analysis of methodological issues. *Journal of Experimental Education*, 32, 147-157.
- Meara, P. (1992). *EFL vocabulary tests*. Swansea: Centre for Applied Language Studies, University of Wales Swansea.
- Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4, 142-154.
- Meara, P., & Milton, J. (2005). *X_Lex: The Swansea Levels Test* [CD-ROM]. Newbury, UK: Express Publishing.
- Meara, P.M. & Miralpeix, I. (2006). *Y_Lex: the Swansea Advanced Vocabulary Levels Test. v2.05*. Swansea: Lognostics.
- Milton, J., & Hopkins, N. (2005). *Aural_Lex*. Swansea: Centre for Applied Language Studies, University of Wales Swansea.
- Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, 63, 127-147.
- Milton, J., & Riordan, O.L. (2007). Level and script effects in the phonological and orthographic vocabulary size of Arabic and Farsi speakers. In P. Davidson, C. Coombe, D. Lloyd, & D. Palfreyman (Eds.), *Teaching and learning vocabulary in another language*. Dubai: TESOL Arabia, pp. 122-133.
- Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, 23, 73-98.
- Nation, P. (1993). Using dictionaries to estimate vocabulary size: Essential but rarely followed procedures. *Language Testing*, 10, 27-40.
- Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, P., & Gu, P.Y. (2007). *Focus on vocabulary*. Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.
- Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press, pp. 6-19.
- Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, 18, 161-175.

- Paribakht, T.S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition*. Cambridge: Cambridge University Press, pp. 174-200.
- Qian, D.D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge. *Canadian Modern Language Review*, 56, 282-307.
- Qian, D.D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52, 513-536.
- Qian, D.D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21, 28-52.
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10, 355-371.
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment*. Mahwah, NJ: Erlbaum, pp. 41-60.
- Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
- Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing*. Amsterdam: Benjamins, pp. 209-227.
- Read, J. (2007). *Extending the Yes/No format as a measure of vocabulary knowledge*. Paper given at the Annual Congress of the Applied Linguistics Association of Australia (ALAA), Wollongong, July.
- Read, J., & Chapelle, C.A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18, 1-32.
- Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, 48, 281-317.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55-88.
- Schoonen, R., & Verhallen, M. (in press). The assessment of deep word knowledge in young first and second language learners. To appear in *Language Testing*.
- Scott, M. (2004). *WordSmith tools. Version 4.0*. Oxford: Oxford University Press.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22, 217-234.

- Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, 12, 309-323.
- West, M. (1953). *A general service list of English words*. London: Longmans Green.
- Yang, H. (1986). A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing*, 1, 93-103.
- Zareva, A, Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, 27, 567-595.
- Zechmeister, E.B., Chronis, A.M., Cull, W.L., D'Anna, C.A., & Healy, N.A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior*, 27, 201-212.