2 *Lexical profiles, learning styles and the construct validity of lexical size tests*

James Milton

Introduction

This chapter will consider in more detail the first of the validity questions which Nation raises in the opening chapter: that of the individual variables each learner will bring to the testing process. Lexical knowledge, like all language knowledge, is not a directly accessible quality like a person's height or weight. In tests, therefore, we rely on the learners themselves to demonstrate their knowledge so we can assess it or measure it. In the opening chapter Nation points out that this is inherently problematic for the validity of a test and its results. If a learner is uninterested and does not try, or guesses a lot, or gives up half way through the test, then the score cannot accurately reflect the learner's true knowledge or ability. The validity of any test of this kind relies on the assumption that learners will behave reasonably, and reasonably consistently, in trying to show what knowledge they have.

In reality we know that learners faced with a test do not always behave either reasonably or consistently. Vocabulary size testing, which makes extensive use of objective style questions, is particularly open to learners using, or attempting to use, test-taking strategies in the hope of maximising their score rather than accurately reflecting their knowledge. A test such as the Eurocentre's Vocabulary Size Test (Meara and Jones, 1990) makes a calculation of a testee's guesswork based on responses to false words contained in the test and, if guessing is sufficiently high, indicates an accurate assessment cannot be made. It is clear from this test that individuals and even groups can behave differently from each other. High guessing levels have been reported among learners who are native Arabic (Al-Hazemi, 1993) and Dutch (Eyckmans et al., in this volume) speakers. By contrast, the Japanese speaking learners reported by Shillaw (1999) display remarkably little guesswork. While tests in this genre attempt to compensate for guesswork, there is no question that attitudinal

factors can be a problem, even to the point of invalidating the results of the test. Eyckmans et al. consider attitudinal factors and guesswork in lexical tests in more detail in the next chapter.

But guesswork and learner attitude are not the only ways in which the qualities a learner brings to this test may affect the measure obtained from it. Learners can vary in other ways and these can also, at least potentially, affect the reliability and validity of the tests we currently use. This chapter considers individual differences in language learning aptitude or learning style and the impact these may have on the validity of the tests of lexical knowledge. The reason why these factors are relevant may not be immediately obvious, but the tests which measure vocabulary knowledge assume that vocabulary is learned in a particular way and it is possible that this is not the case for all learners. This chapter, therefore, will consider whether the frequency model which underlies this sort of test is an appropriate model for assessing vocabulary size in every case, which would in turn challenge the construct validity of the test.

The frequency model of lexical learning

So what is the model so many vocabulary tests are based on? It is a commonly accepted truth in foreign language learning that the more frequent a word is in a language then the more easily, and the earlier, it is likely to be learned. This idea can be traced back at least as far as Palmer who wrote in 1917 that 'the more frequently used words will be the more easily learnt' (1917: 123). Later writers accept this without demur, for example, both Mackey (1965) and McCarthy (1990) repeat Palmer's assertion without reservation. One of the advantages of this idea is that it can be turned into a model which can then be tested empirically. Meara (1992a) does this by graphing up the relationship which, he suggests, should look like Figure 1.

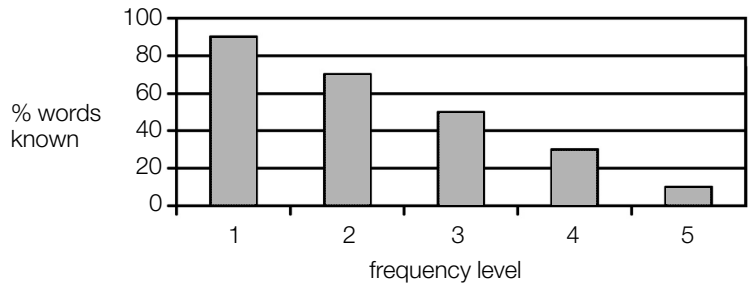

Figure 1 Vocabulary knowledge of a typical learner (Meara, 1992a: 4)

Column 1 represents knowledge of the first thousand most frequent words in a language, column 2 the next most frequent 1,000 words, and so on. A typical learner's knowledge is high in the frequent columns and lower in the less frequent columns giving a distinctive downwards slope from left to right. As learner knowledge increases, this profile moves upwards until it hits a ceiling at 100% when the profile ought to flatten at the most frequent levels and the downwards slope, left to right, shifts to the right into less frequent vocabulary bands.

It is on the basis of this sort of analysis that vocabulary knowledge tests are made. The percentage of words known at each frequency level allows an extrapolation to be made and a calculation of overall lexical knowledge in the foreign language being tested. This is exactly how tests such as the Eurocentre's Vocabulary Size Test (Meara and Jones, 1990) and *X-Lex* (Meara and Milton, 2003b) are constructed. Predictions can even be made of knowledge in frequency levels not tested. Nation's Vocabulary Levels Test (Nation, 1983) tests the 2,000, 3,000, 5,000 and 10,000 word frequency ranges in order to estimate overall lexical competence, confident in the assumption that the frequency levels in between those tested will perform predictably. The tests produced in this way are surprisingly robust. However, there are a number of caveats which need to be acknowledged with this kind of analysis.

One is that frequency information drawn from a wide variety of native speaker sources may not be relevant to foreign language learners who are not exposed to this sort of language but have only textbooks to draw on. Course books will necessarily have to be selective as to the lexis and structures used and lexis in particular is likely to be selected thematically rather than on the basis of frequency. Lexical exposure, particularly at the outset of learning, ought to be different from that which a native speaker might get from newspapers, books and so on. A study of the lexical content of course books reported in Milton and Vassiliu (2000) notes the very high volumes of infrequent vocabulary they include. In principle this might affect the usefulness of frequency-based tests. The evidence on these matters is slim. Meara and Jones (1990), for example, observe that their vocabulary size test is probably not reliable with low-level learners, and while this could be a sampling problem, it might equally well be that the standard frequency models do not reflect the vocabulary which beginners have been exposed to and have learned. But they are unspecific about the point at which it does become reliable. The most recent lexical size tests address this problem in their construction. Meara and Milton's (2003a,b) *X-Lex* Swansea

Levels Test draws on both Nation's (1984) general frequency materials, but also Hindmarsh's (1980) lists, which are more explicitly tied to the vocabulary of EFL textbooks and exams. Nation's Levels Test (1983) takes the same approach.

A second potential problem with the frequency model is that frequency is not the only factor which can influence whether words are learned. Part of speech may affect learning; nouns are usually learned more easily than verbs, which are more readily learned than adverbs. More concrete and imageable words are learned more easily than abstract words. Words which are similar to, borrowed from, or cognate to words in the first language tend to be easier to learn than those which are not. In principle, these other factors ought to affect the slope of this profile. If these other factors have little influence on learnability and the effect of frequency is very strong then the slope of the profile should be steep. Actually, it should be very steep on the left hand side since the most frequent words in a language tend to be very much more frequent than most other words. On the other hand, if frequency of occurrence is not a strong factor affecting learning, because it is overwhelmed by other factors, the slope of the profile should be shallow.

This relationship between frequency and learnability appears to be so self-evident that it is difficult to find a clear empirical demonstration of it in the literature. However, it is not hard to illustrate, at least for populations of learners, and to draw up a lexical profile reflecting the learners' lexical knowledge. Such a profile, drawn from all 227 learners at a school in Greece, ranging in ability from beginners to Cambridge First Certificate level, and created using *X-Lex* (Meara and Milton, 2003a) is shown in Figure 2. The mean score for each frequency level is shown and the resultant graph is remarkably similar to Meara's model in Figure 1. The expected slope from left to right exists demonstrating that the group, as a whole, has a greater know-

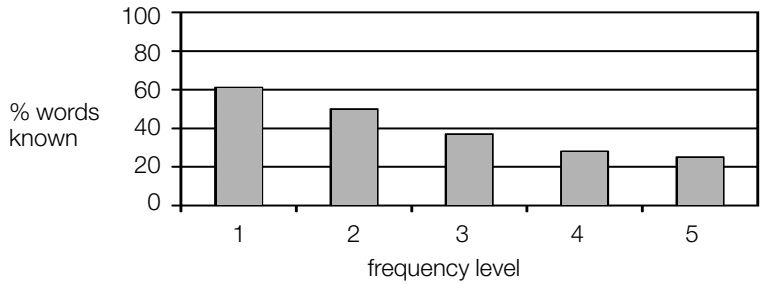

Figure 2 Mean scores for frequency bands

ledge of each succeeding band of greater frequency. It might be argued on the basis of this graph that the profile, which is not a straight line, is steeper to the left between bands 1 and 2 and flattens on the right between bands 4 and 5. This ought to indicate the salience of frequency of occurrence as an influence on the learnability of words. The more frequent a word is, the more likely it is to be learned, as a general rule, and other factors such as the part of speech or concreteness of the words, or the idiosyncrasies of the textbook, do not seem to reverse this trend. A Friedman Test on all of the results confirms the impression that the overall trend is very strong indeed in a population as a whole ($\chi^2 = 512.55$, asympt sig = .000). Very similar results and conclusions have been found among French foreign language learners and are reported by Richards and Malvern (this volume).

But languages are not learned by populations of course, they are learned by individuals. There are good reasons for thinking that individuals may not behave with the same ordered regularity that populations display. Some of the reasons for thinking that individuals may vary are based on the observation of individual profiles. A small study by Vassiliu (1994, reported in Milton and Vassiliu, 2000) notes a dip in some learners' profiles in the second thousand-word frequency band. This is tentatively attributed to a corresponding dip in level two vocabulary presented in the course books his learners used, and he called this feature *level two deficit*. The significance of this is that a test such as the Vocabulary Levels Test draws heavily on level two knowledge which, it seems, may give a misleading impression of overall ability in at least some learners. Subsequent work has suggested that there is indeed something very odd about the lexis particularly in the second thousand-word frequency band. A level two deficit profile is shown in Figure 3.

Meara and Milton (2003b: 5) note a more radical departure from the normal frequency-based profile. Some learners are observed with

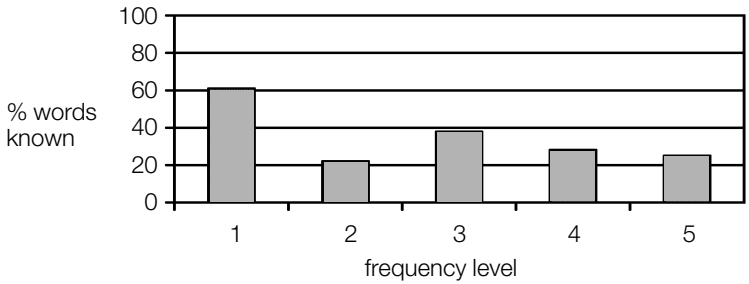

Figure 3 Level two deficit profile

good knowledge of infrequent words but a marked deficiency in knowledge of the highly frequent, structure words which are necessary to put them together to communicate. This produces a profile which is much lower on the left than would be the case in a normal profile and an example is shown in Figure 4. Meara and Milton call this sort of profile *structural deficit*. The significance of this is that a test such as the Eurocentres Vocabulary Size Test is auto-adaptive and relies on a small, initial sample of the most frequent lexis. If scores are low here it presumes that the learner knows even less of the infrequent lexis and does not test it. Such a test would appear likely to underestimate learners with profiles of this sort.

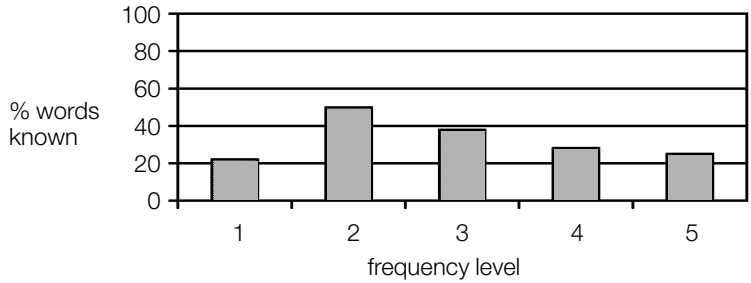

Figure 4 Structural deficit profile

Other reasons for thinking that individual profiles may vary are based on theory. Meara, Milton and Lorenzo-Dus (2001) suggest that learning style might influence the profile, with analytic learners able to acquire structure words (which also tend to be most frequent) easily while memory-based learners will find lexical vocabulary (which is less frequent but tends to be more concrete and imageable) more readily learnable. It might be reasoned that analytic learners should display normal profiles, where structural vocabulary predominates, while memorisers might display level two or structural deficit profiles.

This sort of approach might help explain an anomaly which Vassiliu could not account for. While some of his learners appeared to follow the content of the textbooks and acquire vocabulary with a level two deficit profile, other learners in the same class made up the deficiency in level two lexis and emerged with a normal profile. How could they do this if they were not exposed to the vocabulary? If these learners were strong analytically they might be expected to apply their rule-based systems in generating, bottom-up, their ideas in a foreign language. This approach would inevitably reveal gaps in

knowledge, such as that in level two lexis, and these could be addressed by asking a teacher or looking in a dictionary. Such learners are giving themselves much more opportunity for learning outside the textbook. While this is a nice idea, we have no evidence to suggest whether this really is the case. And in the same way we really have no idea whether all learners behave the same way in the lexis they learn or whether they vary according to aptitude, learning style or level. But this question strikes at the heart of the construct validity of vocabulary size and level tests. If some students do not follow the frequency model of lexical learning then the tests based on this model may make poor estimates of their knowledge.

Frequency profiles and learner aptitude

In order to investigate this for this chapter I have examined the individual profiles generated by the 227 Greek learners described in Figure 2. For the purposes of categorisation I divided the learners according to their profiles as follows:

Normal profile	1 > 2 > 3
Level two deficit	1 > 2 < 3
Structural deficit	1 < 2 > 3

Approximately 60% of learners displayed normal, frequency-based profiles, a further 25% level two deficit (L2D) profiles and approximately 10% structural deficit (SD) profiles. A very small proportion of the results defied classification by these rules. A breakdown of these results over the seven classes involved is shown in Figure 5.

The proportions of each type appear relatively stable over the levels and only appear to change in the two final classes and in particular in the FCE class. Almost certainly, this is the result of ceiling effects. Learners' knowledge of individual lexical levels appears to peak at

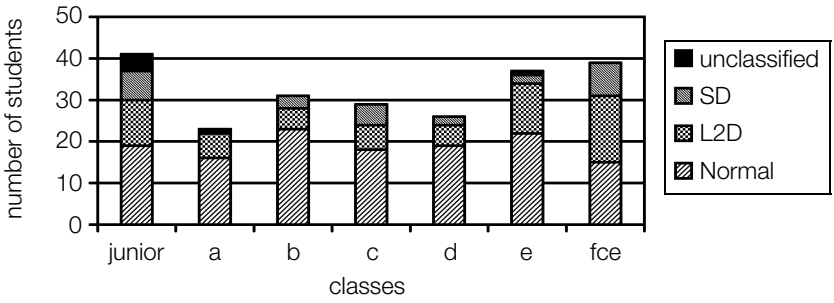

Figure 5 Distribution of profile types

around 85–90%, and not at 100% as expected, and in the highest class this peak has been reached in the most frequent bands. At this stage, a shift of a single mark can change the profile, something that does not happen in lower classes where the differences between frequency band scores are greater.

It might reasonably be questioned whether these profiles are stable, and therefore a reflection of some characteristic of the learner's vocabulary knowledge, or whether they are a result of some variation which the testing method generates. Reliability measures of vocabulary size and level tests (for example, by Adamopoulou, 2000) generally suggest that they are incredibly reliable by language testing standards. Test–retest correlations of 0.99 suggest that the profiles are unlikely to change. With the Greek learners in this particular study, 29 learners took the test twice (a different form each time) and the profiles produced were compared. The results showed that in each test there were 15 learners with normal profiles and 14 learners with level two deficit, but that there was some movement between these groups and this is shown in Figure 6.

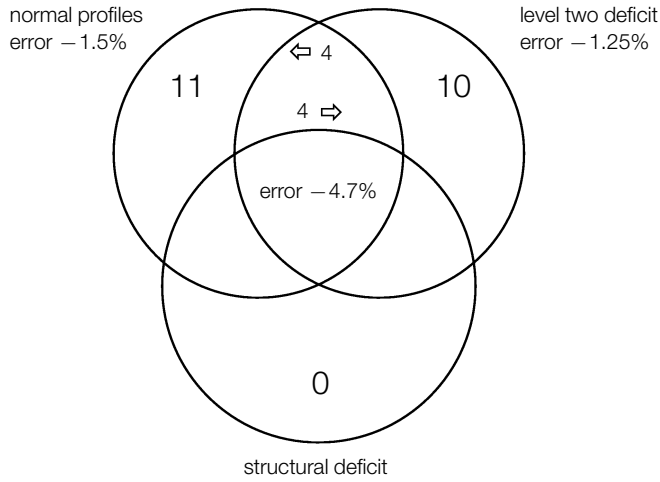

Figure 6 Movement between profiles

The majority of learners, 21 out of 29, retained stable profiles between tests. Eight learners, however, changed profile, four moving each way between normal and level two deficit profiles. The reason for profiles destabilising may lie in the guesswork which learners use, calculated as error figures (the percentage of false words identified as real words) which are also shown in Figure 6. Learners whose profiles

remain consistent show very little error, approaching an average of only 1%, and must be very sure of their vocabulary knowledge since they guess hardly at all. Learners whose profiles change have much higher error rates, three or four times higher, and these errors, or rather the guesswork that produces them, may well be enough to destabilise the profiles. This observation lends additional weight, if it were needed, to this chapter's opening point that guesswork can seriously destabilise a test's results.

The first tentative conclusion to be drawn from this is that it seems possible that as many as a third of learners may depart in some way from the frequency model of vocabulary learning. Despite the strength of the frequency effect on learners as a population, it appears that there is some systematic variation among learners as individuals. In principle, this should challenge the validity of frequency-based lexical size tests.

It might be expected, if these different profiles are the product of the learners' varying aptitude, and in particular memory and analytic skills, that learners will display different scores on aptitude tests designed to evaluate just these qualities. If the theories of Meara, Milton and Lorenzo-Dus (2001) are correct then those with a normal profile should do comparatively well on tests of analytic ability while those with level two deficit profiles should score comparatively well on tests of memory. The 21 Greek learners with stable profiles were therefore also asked to take two tests from the Meara, Milton and Lorenzo-Dus (2001) range of aptitude tests. These were LAT_B, a paired associates learning task designed to test memory in language learning, and LAT_C, a language rule recognition task designed to test inductive and analytic language learning skills. The learners were grouped according to their profiles, eleven normal profiles and ten level two deficit, and their scores on these aptitude tests calculated. Mean scores are presented in Figure 7.

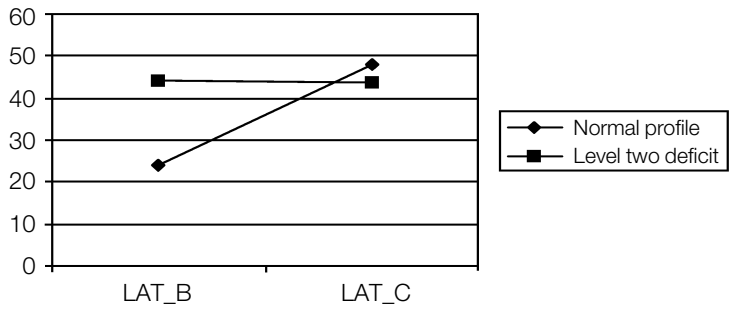

Figure 7 Mean scores on aptitude tests

Broadly, these results bear out Meara, Milton and Lorenzo-Dus' (2001) expectations. Learners with normal profiles score higher on the analytic skills test, LAT_C, than the learners with level two deficit, while learners with level two deficit score higher on the memory test, LAT_B, than those with normal profiles. The LAT range of aptitude tests have been normalised so it appears that the normal profile learners are much stronger in analytic skill than in memory, and this would support the idea that they are likely to be analytically orientated learners. The difference in their scores, and the difference compared to the level two deficit learners is very marked. The level two deficit learners score higher in memory than in analysis but the difference is much smaller and given that normalisation was not carried out on this age group it would probably be a mistake to read too much into the similarity of this group's aptitude scores. An ANOVA analysis of the results shows that there is a group effect and a test effect and, more importantly, a significant interaction between group and test. Even though these groups are very small, the results suggest there is a difference in aptitude between learners with the two different profiles. Results are shown in Figure 8.

Source	LATBC	Type III	df	Mean Square	F.	Sig.
LATBC	Linear	1248.200	1	1248.200	6.981	.018
LATBC*Style	Linear	1355.756	1	1355.756	7.583	.014
Error (LATBC)	Linear	2860.800	26	178.800		

Figure 8 Test statistics on aptitude test scores: within subjects contrasts

These results suggest that, as Meara, Milton and Lorenzo-Dus (2001) indicated, different learning strengths and styles really can influence the foreign language lexis that learners acquire in class. The frequency effect may not disappear from their profiles entirely, but learners may not always learn the vocabulary in the first two 1,000-word frequency bands with the ease and facility which the frequency model suggests they should. In these ranges, a learner's aptitude or style may help determine what is learned.

Conclusions

What then can be concluded from this, and how important to the validity of the vocabulary tests we use, is this observation concerning the effect of aptitude on lexical learning? To put this into perspective, one thing that should emerge from this chapter is that the frequency

model appears to be a really very cogent model of learning as a whole. Notwithstanding these observations of individual variation it cannot be discarded. The frequency effect on vocabulary learning is very strong and this should not be lost. The individual variations observed here appear to affect only the most frequent 2,000 words and normal interaction between frequency and learning appears to assert itself again beyond these levels. On the face of it, therefore, vocabulary size and knowledge tests based on frequency retain very good construct validity. In this respect the frequency model they are based on is probably better than the models underlying many other widely used language tests.

Once this is said, however, the evidence, though still slight, supports the idea that individuals may vary in the vocabulary they learn according to learning style or particular aptitude strengths, and that this produces different frequency profiles. Analysis of Greek learners has supported the theoretical assumption made by Meara, Milton and Lorenzo-Dus (2001) that in addition to the normal frequency profile two other types of profile are identified. A level two deficit profile where the 2,000-word level is disproportionately low in comparison to the others, and a structural deficit profile where the 1,000-word level is disproportionately low. Students with level two deficit profiles appear strong in memory compared with students with normal profiles, who are stronger in analysis. About three or four in every ten students tested displayed these odd profiles.

This is an interesting finding but what is the implication for the validity of lexical size tests based on frequency? The remainder of the frequency profile remains generally frequency based so the impact on lexical size tests need not be so great if the methodology is not dependent on knowledge of the 2,000 most frequent words. Meara and Milton's (2003b) *X-Lex*, which tests each of the first five 1,000-word frequency bands separately, and then provides a score out of the 5,000 most frequent words, would appear unaffected. Its construct validity appears unchallenged by these findings.

Other tests are likely to be more affected. The Vocabulary Levels Test (Nation, 1983) relies for its estimation of ability quite heavily on knowledge of the second 1,000-word frequency band. Overall ability is inferred, in part, from this knowledge. Clearly learners with level two deficit are likely to perform disproportionately badly since an inference will be made that other levels will be low when this need not necessarily be the case. Other levels are tested, of course, and this will mitigate the effect of the underestimation. But it is a concern where as many as one in four may have profiles of this type.

Potentially the test most likely to be affected is the auto-adaptive

test which Meara and Jones (1990) created for Eurocentres. This test makes an initial estimate of overall vocabulary size based on knowledge of the most frequent vocabulary levels, before testing in depth at the language level the initial test elements suggests. Where learners have disproportionately low knowledge of this frequent vocabulary, the later in-depth test is likely to be at the wrong level. The conclusion to be drawn from this chapter is that these frequent levels may not be the best predictors of overall lexical size and may underestimate in as many as one in three cases. Meara and Jones recognise that this is likely to be a particular problem with low-level learners and warn against this, but it is a concern.