Some notes on the Eurocentres vocabulary tests.

Paul Meara
***Birkbeck College, London University*.**

The paper that I gave at AFinLA described a computerised vocabulary test that we have developed at Birkbeck College, as part of our ongoing research on vocabulary skills in an L2. These so-called yes/no tests have been described more fully elsewhere (Meara and Jones 1988, and Meara and Buxton 1987). In this report, then, I will not go over this material again. Instead, I will pick up three points that were raised in the discussion at the AFinLA meeting.

Firstly, however, a brief outline of the tests is necessary for readers who are not familiar with them. The tests are known as the *Eurocentres Vocabulary Tests,* since they were originally developed in response to a commission from Eurocentres, a large group of language schools based in Switzerland. Like many other language schools, the British part of Eurocentres runs a series of short, four-week intensive courses aimed at non-native speakers in English. This rapid turn round means that the schools are faced with a major administrative problem at the start of each course: how to grade the new students and place them in an appropriate class. Traditionally, Eurocentres has done this by means of their *Joint Entrance Test* (JET), a specially developed test package, consisting of a grammar test, an auditory comprehension test, a reading comprehension test and an oral interview. JET takes about one and a half hours to administer, and has to be marked manually. This involves a significant amount of staff time.

The test we developed for Eurocentres does the same job as the JET test: that is, it is designed to work as a placement test, grading students roughly according to their overall ability in English as a foreign language. It is, however, radically different from the more traditional JET test format. Our test is simply a vocabulary test. It makes an estimate of the overall vocabulary size of each students, and places the students in a class which is made up of people with similar overall vocabulary scores. The tests as we developed them are fully computerised, take approximately 10 minutes to run, and score themselves automatically. The class groupings produced by the vocabulary test are very similar to the groupings produced by the JET test. In general, scores on the JET test correlate fairly well with the scores on the vocabulary test (generally in the region of

.6 to .8 for a group of 100 plus testees). Some anomalies to arise, but such anomalies are generally few in number.

How is this miracle of efficiency achieved? The basic idea behind the test is extremely simple. Learners are presented with a series of words, one that time, and are asked to indicate by pressing one of two keys on the computer whether they think they know that were not. The "words" actually consist of two types of items: some of these items are genuine words, the others (about a third) are non-existent words which the testees cannot possibly know. An example of the sort of items the testee sees is shown in table 1 below.

Table 1 a sample vocabulary test.

block a

adviser	ghastly	contord	implore
morlorn	patiful	profess	stourge
moisten	discard	disdain	gleanse
weekend	boyralty	partine	indoors
storage	vibrade	dostage	refusal
sarsage	bariner	mertion	smother

block b

mascule	palangane	bezel	maparotomy
peneplain	rangue	aliver	orduad
leat	prunella	gamelkind	masquinade
ablegate	mittimus	rickwall	quoddity
algorism	myosote	killick	windlestraw

The words in table 1 are divided into two blocks. The first block consists of a set of highly frequent English words, together with some non-words. The second block consist of real words which are very infrequent, together with some non-words. Most native speakers of English, and most high-level learners, find it very easy to decide which are the real words in the first block. The second block is very much harder, and even native speakers of English have a great deal of difficulty in distinguishing which of these items are real words, and which are imaginary ones. The real words in the first block are highly frequent items that every native speaker would be expected to know. The real words in the second block are very low-frequency items. The Eurocentres

vocabulary test uses this basic principle. The test is divided up into a number of blocks. The first block contains a sample of items from the first 1000 most frequent words in English; the second block is a representative sample from the second 1000 most frequent words in English; and so on up to block 10, which comprises a representative sample from the 10th 1000 most frequent words in English. The test samples each block in turn, and calculates what proportion of the words in that block the testees knows. The total vocabulary score is calculated by adding together the scores for each block of words.

The scoring system takes account of two types of response made by the testees. The score we are really interested in is the **Hit Rate** (i.e the proportion of real words that the testees think they know), but we also have to take account of the **False Alarm Rate** (i.e. the protportion of imaginary words that the testees think they know). The computer calculates an estimate of the true Hit Rate, by adjusting the actual Hit Rate in the light of the False Alarm Rate. Some examples may help to make this clear (Table 2).

Table 2: some typical combinations of Hit Rats and False Alarm Rates

	Hit rate	**False Alarm Rate**
Case A	**100%**	**0%**
Case B	**50%**	**0%**
Case C	**100%**	**100%**
Case D	**50%**	**50%**
Case E	**60%**	**5%**

Testee A in table 2 claims to know all the real words, and rejects all the non-words. No adjustment is made his score for real words. Testee B in table 2 claims to know 50% of the real words, and rejects all the non-words. Again, no adjustment is made to his score for real words. Testee C claims to know all the real words AND all the non-words. His adjusted score is zero. Testee D claims to know 50% of the real words and also 50% of non-words. His adjusted score is also zero. Testee E, who is more usual than the other types, claims to know 60 percent of the real words, but also claims to know 5% of the imaginary words. This suggests that his actual Hit Rate slightly over-estimates his real Hit Rate, and the actual Hit Rate is therefore adjusted downwards slightly. The mechanism for doing this is based on *Signal Detection Theory* models (cf. Zimmerman, Broder, Shaughnessy and Underwood 1977), but these will not be discussed here.

The advantages of a test of this type will be immediately apparent to anyone who has tried to work with more complex tests. The task required of the students is extremely

simple; the test requires no complex development of items; scoring is straightforward and can be automated; the simplicity of the task means that a very large number of items can be tested in the short space of time; parallel versions of the test can be produced with a minimum of effort; and finally, the test does not appear to have any deleterious wash back effects -- the only way you can do well on it is by knowing lots of words.

Whenever I have talked about these tests to applied linguists, three main questions always seemed to crop up in the discussion, and it is to these that we now turn.

The first comment concerns the level of word knowledge that the tests home in on. In particular, it is generally assumed that the yes/no format of the tests can only assess passive recognition ability. This is clearly okay as far as it goes, but clearly there is a need for more accurate, finely tuned tests which also assess students' abilities to use the words in question.

When this criticism was first raised with me, my immediate reaction was to agree wholeheartedly. Clearly, the yes/no format does test passive recognition skills, in that the testees need only to say whether they recognise the word or not. Indeed, they don't even need to say that they know what the word means: they only need to say YES to words that they know are words in the L2 in order to be able to score.

There are, however, two reasons why this criticism may not be as important as it looks. Firstly, there is nothing intrinsically wrong with measuring passive vocabulary recognition. You can argue (and I did argue) that passive recognition is the basic, rock bottom skill on which all the other skills rest, the *sine qua non*, as it were, of vocabulary skills. A student who does not even recognise that a particular string of letters is a word in the L2 is not likely to be able to do very much with it. Similarly, a student who CAN do something with a word (no matter how much or how little) will certainly be able to recognise that it is indeed a word. Secondly, there is obviously some relationship between active vocabulary size and passive vocabulary size. Quite what this relationship is remains an empirical question, of course, but there must be some limitations on how active and passive vocabulary size can co-vary. Passive vocabulary must obviously be bigger than active vocabulary, and it seems reasonable to assume that people score highly on a test passive vocabulary, would also perform well on an active vocabulary test if one existed. My guess is that active vocabulary size and passive vocabulary size vary systematically, and that under normal circumstances, passive vocabulary size is generally X percent bigger than active vocabulary size. It would certainly be unusual to find a learner with a huge passive vocabulary and a tiny active one, except in very special circumstances, once the initial stages of learning a language

have been passed through.

More recently, however, I have come round to the view that the yes/no test format is probably less passive than it looks at first sight. I reached this conclusion after I had been developing a set of tests for Spanish as a second language. Most English speakers have a large passive recognition vocabulary for Spanish. Many Spanish words are cognate with low-frequency English words, and English speakers who know French or even better Latin, often comment that Spanish is easy to read. However, you can exploit this tendency in Spanish by devising non-words which look as though they are genuine cognates, but actually aren't. So, for example, you can take the stem form TARDE (meaning LATE) and combine it with the affixes RE- and -IMIENTO to form a non-word RETARDIMIENTO. This looks as though it ought to mean something like "the act of being late". (The real Spanish word for this concept is RETRASO.) Now, an English speaker coming across RETARDIMIENTO in a text would probably have no trouble guessing what it means. An English speaker coming across the word in our test is faced with a slightly more subtle problem. He has to decide not if he knows what it MIGHT mean, but if he knows that it really exists in Spanish. In practice, if the testee says YES to items like this, then his False Alarm Rate goes up, and this means that is Real Hit Rate is adjusted downwards. This adjustment means, in effect, the testees are penalised for assuming they know what an item like RETARDIMIENTO means. The final score is thus not a pure reflection of passive knowledge; it is also influenced by how confident the testees are about their ability to use the words they claim to know.

This brings us to the second problem which generally arises in discussion of the tests: the extent to which the scores are affected by the nature of the non-words. The argument here is that non-words with different properties might affect speakers of different language backgrounds to a greater or lesser extent. For instance, take the non-word LOYALMENT in English. For Spanish speakers, this word is likely to cause some problems. There exists a cognate form in Spanish LEAL (meaning *loyal*), which combines with the -MENTE ending to form the adverb LEALMENTE (*loyally*). This correspondence probably makes Spanish speakers more willing to accept a form like LOYALMENT. For German speakers, on the other hand, LOYALMENT raises different problems, all of which are to do with how LOYAL works in English (is it a noun or adjective?). There are two main answers to this objection. Firstly it seems to be the case that testees using this test are very cautious in the number of non-words they accept anyway, and this tendency may be enhanced by careful wording of the instructions. This means that, in practice, the choice of non-words may not really matter very much. Secondly, the work we have done so far suggests that the tests do work slightly differently for testees with different L1s. The way we establish this is to compare the

correlations between our vocabulary test scores and test scores on another different test (e.g. the Eurocentres JET test). We then calculate the regression line for the data. What we are finding is that the slope of this line varies slightly from one language group to another. We also find that the intercept varies from one language group to another. German speakers, for instance, have lower intercepts and steeper slopes than Italian speakers. The differences are not very large, but they do suggest that cognate effects may be operating, and that the distribution of cognates in the L1 and L2 might affect the way the scores come out. This means, however, that the choice of non-words may be relatively unimportant within the overall framework of the test.

The third problem that arises in discussion is also related to the cognate issue. In the early versions of the tests, we initially assumed that learners of English as a second language would be sensitive to the frequency characteristics of the words tested. In particular, we guessed that most learners would have a good grasp of highly frequent words, and the lesser grasp of less frequent words, and that as we moved from frequent to less frequent words, the proportion of words a learner knew would decline gradually. Recently we decided to check this assumption out, by asking our program to print out not just a total for each testee, but also a profile, showing which parts of the total vocabulary a testee does well on, and where they perform poorly. The results of this output was surprising. Many testees do indeed perform as we expected. Figure 1, for instance, shows a high-level Swedish speaker of English, who clearly has no problem with high frequency words in English, and knows most of the vocabulary in the 5-10 thousand words range. Figure 2, however, shows a rather different pattern. Here the testee, a native speaker French, has severe difficulties with simple vocabulary, and is much more at ease with the low-frequency vocabulary. The obvious explanation for this aberrant pattern is that the testee in figure 2 followed a traditional literary based course in English, which has left him totally unequipped for the demands of ordinary intercourse in English. The general point, however, is that for whatever reasons, not all testees fit the neat patterns that we have expected, and some of them show surprising and unpredictable patterns in their profiles. This has serious implications for our interpretation of the overall vocabulary score estimate produced by the computer program. In particular, it suggests that the overall score on its own may be a misleading piece of data. It can only be taken at face value if a profile for a particular subject fits the expected pattern. Where the actual pattern found turns out to be very different, then the overall score can be severely misleading. We are currently working on alternative ways of scoring the test, so that both the profile and the overall total score are taken into account.

Figure 1: Vocabulary profile from a fluent non-native speaker of English. L1=Swedish.

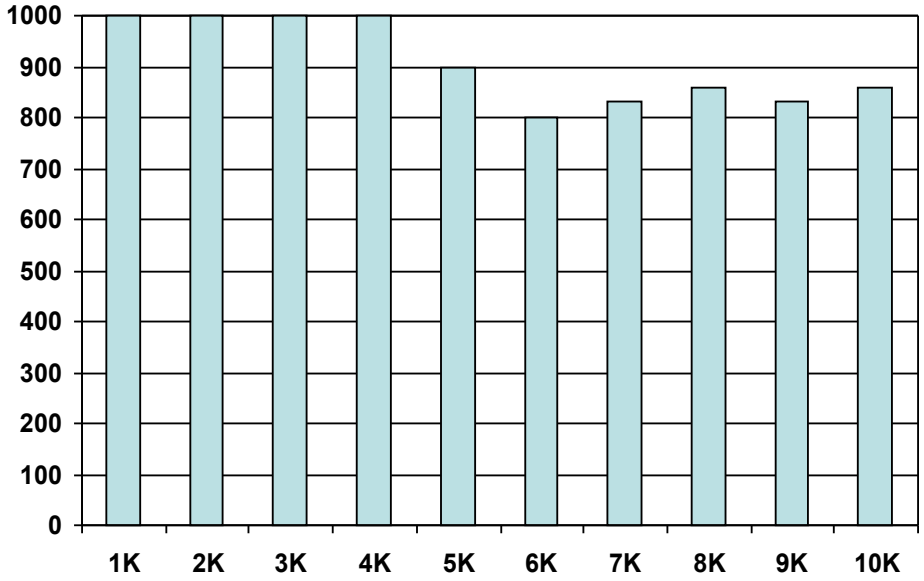
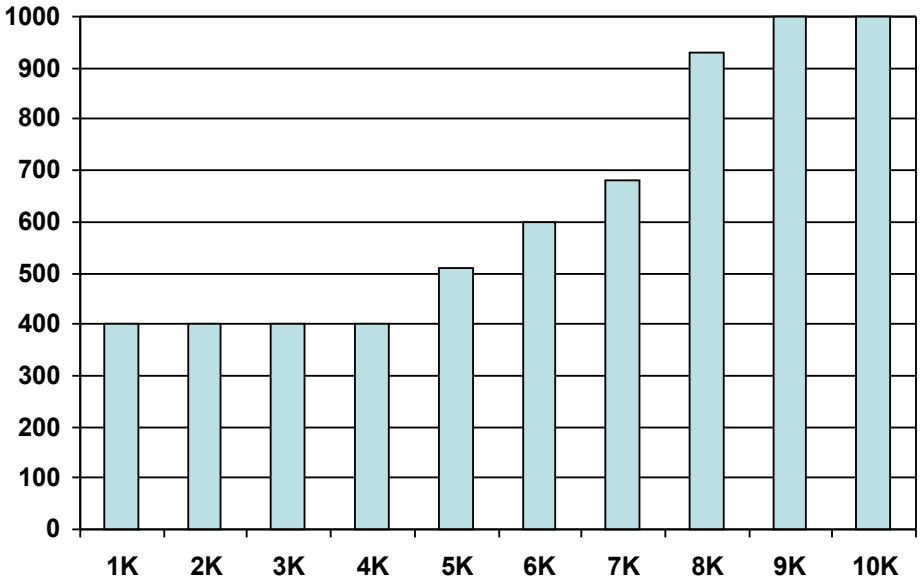


Figure 2: Vocabulary profile of a typical "academic" learner of English. L1=French.



Summary
This paper has briefly outlined the thinking behind the Eurocentres Vocabulary Tests, and looked at three critical comments which came up in discussion of the tests at the AFinLA meeting. Though these criticisms are obviously important ones, none of them seriously invalidates the tests. They do, however, highlight the need for more thorough background research into some of the assumptions that the tests rely on. It is, of course, a terrible cliche to end a paper with a plea for more research. In this case, however, it is clear to me that the Eurocentres test potentially makes it possible to answer a very large range of questions that we previously had no grip on. Further research on its reliability and usefulness, and the limits of its application, might make a real contribution to our understanding of second language acquisition.

References

Meara, PM and B Buxton 1987.
An alternative to multiple choice vocabulary tests. *Language Testing* 4(1987), 142-154.

Meara, PM and G Jones 1988.
Vocabulary size as a placement indicator. In: **P Grunwell** (ed.) *Applied Linguistics in Society*. London: CILT.

Zimmerman, J, P Broder, J Shaughnessy and B Underwood. 1977.
A recognition test of vocabualry using signal detection measures, and some correlates of word and non-word recognition.

This paper first appeared in **J Tommola** (ed.) *Foreign Language Comprehension and Production*. Turku: AFinLa Yearbook. 1990. 103-113.