Applied Linguistics 21/4: 490–516

Testing L2 Vocabulary Knowledge at a High Level: the Case of the *Euralex French Tests*

PAUL BOGAARDS

Leiden University

Testing L2 vocabulary knowledge should take into account all aspects of what is implied by lexical knowledge: knowledge of form, meaning, morphology, syntax, collocations, and use in discourse. Different types of tests will be necessary to tap all these aspects. Some general aspects of vocabulary testing are briefly discussed in this context. In the second part of the paper, a new type of test which aims at testing very high levels of L2 vocabulary knowledge is analysed and investigated. Although the validity of the test as it stands is questionable, it is shown that it is possible to improve on both its reliability and its validity, and to adapt it to the level of vocabulary knowledge which is common with advanced university students.

Testing vocabulary knowledge in a second or foreign language is not as straightforward an affair as is sometimes thought. Depending on what exactly one wants to know about L2 lexical knowledge, one has to select the appropriate materials and adequate procedures to arrive at valid and reliable results. As lexical knowledge comes in very many forms and presents a lot of different aspects, this means that there is not one single valid way to measure L2 vocabulary knowledge. Different types of tests are needed to address different aspects of the lexicon and different formats may be more or less adapted to different levels of vocabulary knowledge and to different types of questions the teacher or the researcher wants to answer.

In the first part of this paper I will briefly discuss the question of what is implied by 'knowing a word' and I will comment on some general problems concerning the design of vocabulary tests. In the second part I will report on a piece of empirical research about one particular form of test, the *Euralex French Tests* (EFT), which was devised to measure very high levels of lexical knowledge in French as a foreign language. I will give an analysis of the content of these tests and I will present experimental studies on their validity, their reliability and their discriminating power, and I will describe some necessary further developments of the tests. In doing so I will try to demonstrate that an EFT-type of test is potentially a valid, reliable, and useful tool for the measurement of specific aspects of lexical knowledge in an L2.

TESTS OF LEXICAL KNOWLEDGE IN L2

In order to know what aspects of lexical knowledge can or should be tested, it is important to have a clear view of the nature of what can be known about lexical units. Several attempts have been made at establishing an exhaustive list of aspects of lexical knowledge. Cronbach (1942) distinguishes five aspects: generalization (knowing the definition), application (knowledge about use), breadth of meaning (knowing different senses of a word), precision of meaning (knowing how to use the word in different situations) and availability (knowing how to use the word productively). Richards (1976) gives a list of seven criteria:

1 knowing the degree of probability of encountering that word in speech or print;
2 knowing the limitations imposed on the use of the word according to variations of function and situation;
3 knowing the syntactic behaviour associated with that word;
4 knowledge of the underlying form of a word and the derivations that can be made from it;
5 knowledge of the network of associations between that word and other words in the language;
6 knowing the semantic value of a word;
7 knowing many of the different meanings associated with the word.

More recently, Nation (1990: 31) has proposed four dimensions of lexical knowledge: form (oral or written), position (grammar and collocations), function (frequency and appropriateness) and meaning (conceptual and associative). All these forms of knowledge can be productive or receptive.

These three approaches to the question what it means to 'know a word' are more or less complementary to each other. For instance, Nation adds aspects concerning the knowledge of the written or spoken form of the word, but he does not mention problems of morphology, which are taken into account only by Richards. Whereas Cronbach does not speak about associations or collocations, Nation seems to ignore the possible polysemy of words.

To different degrees, in these approaches the word is taken as the unit one has knowledge about. It is not specified, however, what is meant by a word. As is well-known, the concept of 'word' has never been very clear in linguistic theory, although many different definitions have been formulated. Generally, such different elements as *finch*, which has a single meaning, and *fine*, which has many and belongs to several parts of speech, *to*, which has above all a grammatical function, and *bye*, which has rather a pragmatic function, *free*, which is a simple form, and *freedom* or *freelancer*, which are different types of complex forms, are all called 'words'. On the other hand, elements like *give up*, *heat wave* or *out with*, which may be considered to have the same kind of semantic and grammatical behaviour as the elements called 'words' are mostly excluded from that category.

Following Cruse (1986), I prefer to use the concept of *lexical unit*, which is assumed to be the union of one stable meaning and a well-defined form. According to this definition, there are, on the one hand, different lexical units sharing the same form: *party* (political party, my neighbour's party, rescue party, etc.). On the other hand, elements like *give up*, *heat wave* and *out with* as well as *spill the beans* or *Fortune favours fools* are considered as lexical units in their own right. As I have tried to make clear elsewhere (Bogaards 1994, 1996, forthcoming), L2 learners may be assumed to learn lexical units, not only 'words'.

As to the aspects that may be learned about lexical units in an L2, at least the following can be listed:

1 **form**—Learners have to get acquainted with the written and/or the spoken form of the unit. Knowing that a given form does indeed belong to a given language seems to be a first stage of knowledge.
2 **meaning**—Knowledge of the semantic side of a lexical unit may come in different shapes. One can have a vague notion, e.g. that *haematin* has something to do with the blood or that a *beech* is some kind of tree. It will be clear that knowing the differences between *arrogant*, *presumptuous*, and *superior* is of another order. In other words, knowing something about the meaning of a lexical unit does not necessarily mean that one knows its meaning nor does it imply that that element has been fully integrated into the semantic network it belongs to or that one has understood all its connotations. Moreover, knowing (something of) one meaning that is associated with some form does not imply knowledge of other meanings that that same form may have, as in the case of *party*.
3 **morphology**—Lexical units have their own conditions on derivation and compounding. *Gracefully* and *graciously* have relationships to two different lexical units which share the form *grace*. Even when the morphological mechanisms of the L2 are well understood, many of the actual relationships have to be learned one by one (see also Bogaards 1994: 53–7). Especially for productive use it is difficult for the L2 learner to know whether a given form is possible and in what sense it may be used.
4 **syntax**—A learner who knows the rules of the syntax may make many mistakes by not applying the right rules to the right lexical units. This applies especially to verbs and, to a lesser degree, to adjectives. Learners have to find out how many and what types of arguments are obligatory or possible with a verb in a given sense, or which prepositions have to be used with a verb or an adjective in some specific sense, i.e. with a particular lexical unit.
5 **collocates**—Whereas some lexical units, like *very* or *red*, seem to be usable in combination with a great number of other elements of a given category, others have a very restricted realm of use. Most of the collocations that seem so natural to native speakers make for great trouble for L2 learners.

6 **discourse**—In what type of discourse and to what effect can lexical units like *furthermore*, *moreover* and *what is more* be used? Which lexical units are to be avoided when speaking to someone belonging to another ethnic group or when writing a letter of application? Knowledge of style, register and appropriateness of particular senses of a same word is notoriously difficult to acquire for L2 learners, and it takes a long time before they have a feeling about the relative frequency of lexical units in different kinds of contexts.

This list may make it clear that there is much more to testing vocabulary knowledge than just ascertaining that learners know some words. Lexical knowledge is more a matter of degree than of 'yes' or 'no' and implies far more than just knowing one meaning for each form (Wesche and Paribakht 1996: 14). In addition, all these types of knowledge may be readily accessible and available at any moment for productive use or else be restricted to certain contexts or only recognized when used by others.

In the last twenty years, several studies have been devoted to the question of how vocabulary knowledge can or should be measured. In most cases a new test format is discussed and compared to more traditional ones. Such is the case with the *Vocabulary Levels Test* (Nation 1983), the yes/no-test (Meara and Buxton 1987), the checklist (Read 1988), the C-test (Singleton and Little 1991, Chapelle 1994), the *Lexical Frequency Profile* (Laufer and Nation 1995) or the *Vocabulary-size Test of Controlled Productive Ability* (Laufer and Nation 1999). Although more general remarks on vocabulary testing are not absent from these studies, more principled discussions about the ways and techniques that can be used to gain insight in the lexical knowledge of L2 learners are harder to find.

As an introduction to his *Word Associates Test,* Read (1993: 355) proposes 'four somewhat overlapping dimensions that can be seen to underlie the design of vocabulary test formats'. According to Read, when deciding on a test form, one has to take into account the following aspects:

1 simple or more complex test formats, the yes/no-test being the simplest form;
2 verifiable responses or self-report (i.e. testees have to say whether they know the word or not);
3 size or quality of knowledge, or what is called breadth as opposed to depth of lexical knowledge;
4 testing in context or in isolation.

As can be seen, the first two of Read's dimensions have to do with the type of reaction that is required from the test taker, whereas the third one is first of all concerned with the type of knowledge that is measured and the fourth one is about test conditions.

In order to describe the existing test forms which are used for measuring breadth of vocabulary knowledge and as an introduction to their *Vocabulary*

Knowledge Scale, Wesche and Paribakht (1996: 15) present the following list of characteristics:

1 how target vocabulary items are selected;
2 the nature of the task presented to the testee;
3 the test response format used;
4 criteria for judging open-ended responses.

In this case the second and the third characteristics seem to overlap with the first two dimensions of Read, whereas the first characteristic adds an important element of test development and the last one takes into account an aspect of test evaluation.

Joe, Nation, and Newton (1998 unpublished) describe eight different test formats by combining three factors affecting test item difficulty. These factors are:

1 the receptive/productive (or passive/active) distinction;
2 the recognition/recall distinction (roughly the distinction between multiple choice items and fill in items);
3 the imprecise/precise distinction.

In each of these distinctions, the first element is said to be easier than the one after the slash. Among these distinctions, the second one is again about type of response, whereas the third factor, as it has to do with the 'degree of accuracy required in the answer', seems to have some overlap with the last characteristic of Wesche and Paribakht. As to the first distinction, Joe *et al.* is the only one of these three attempts at describing important aspects of vocabulary tests to take into account the difference between receptive and productive knowledge.

What these three lists have in common, in spite of their quite different aims, is that they try to present rather heterogeneous aspects of testing in one framework: aspects of content, format, construction, and/or evaluation of vocabulary tests tend to be amalgamated. Although these aspects are indeed very much interrelated, it is necessary to keep them conceptually apart, and all the more so in view of the very complex nature of lexical knowledge.

From the discussions which the authors mentioned so far devote to existing and new types of vocabulary tests, it appears that tests measuring vocabulary size or breadth of lexical knowledge are the most widespread. Wesche and Paribakht (1996: 26) list twelve different formats that can be used to evaluate this aspect and they conclude that these tests 'are useful in providing broad characterisations or comparisons of vocabulary size, such as for admission to programs requiring knowledge of the given language or placement in language courses'. That is to say that for practical goals, existing measures of vocabulary size seem acceptable. But at the same time, Meara (1996: 38) states that one of the basic problems in vocabulary research is 'that there are no reliable tests of vocabulary size'. On the other hand Wesche and Paribakht (1996: 26) rightly stress that for L2 users other types of tests are also necessary

because partial knowledge is important and because 'advanced learners need depth and speed of access as well as range in their vocabulary knowledge, for ease and precision of comprehension as well as for effective composition and oral expression'.

This brings us to an aspect that is only seldom taken explicitly into account when vocabulary testing is discussed, although testers know that it is the first thing one should be aware of (Hughes 1989: 7, 48): (vocabulary) tests are used for rather different purposes. On several occasions it has been suggested that tests of lexical knowledge only answer one question, which is presented as the only one worth answering. Meara (1996: 37), for instance, states that '[t]he basic dimension of lexical competence is size' (see also Read 1988: 14). However, as was stated above, Wesche and Paribakht are not only interested in global measures of vocabulary size, but also in partial knowledge and in depth of knowledge. Recently Read (1997: 311–17) has made an attempt at classifying vocabulary tests according to purpose. Next to forms of integrative lexical measures such as the cloze procedure and the C-test, he tries to make a distinction between vocabulary measures used in a research context or in a school setting. Unfortunately, he seems to mix up these two categories when he mentions Meara's yes/no-test and Nation's *Vocabulary Levels Test* as useful instruments for classroom teachers and school administrators in the section about research.

For several reasons, then, it is important to think about the measuring of other aspects of lexical knowledge. Even if one accepts size or breadth as an essential dimension, this does not imply that other dimensions are unimportant. Especially for advanced learners, it is important that they acquire more senses of polysemous words and learn more about possible collocates, special uses, and so on. Teachers and researchers should be able to investigate progress in vocabulary acquisition, even in situations where no or very few new word forms have been learned.

Consistent with this view, Read (1988: 17) argued that particularly for achievement testing 'we need to concentrate on depth of knowledge: how well are smaller sets of key vocabulary items known?' Later on he developed the *Word Associates Test*, a format that involves 'a simple response task and allow[s] broad coverage of a set of words while, at the same time, probing depth of knowledge of words in some meaningful way' (Read 1993: 358). In this test, items are composed of a target word and eight other words, four of which have some sort of relationship with the target word and four do not. Testees have to select the four words having a relationship. For a word like *edit*, related words would be *revise* (paradigmatic relationship), *film* (syntagmatic relationship) and *publishing* (analytic relationship, i.e. a relationship where the related word 'represents one aspect, or component, of the meaning of the stimulus and is likely to form part of its dictionary definition'). After a pre-test, Read has given two tests of 50 items each to more than 200 learners of English. The tests turned out to be very reliable (KR-20 of .92), the intercorrelation between the two tests was very high (.97) and the

correlations with a test of words taken from the University Word List were satisfactory (about .75). However, it was found that the associates of some of the stimuli had semantic connections among themselves and verbal reports of some test-takers made the author 'cautious about interpreting the successful identification of two or even three associates in an item as necessarily evidence that the test-taker knows the stimulus word. It seems that, frequently, the ability to select correct responses is based on a combination of good vocabulary knowledge and a certain resourcefulness in seeking possible associates, as well as the confidence to make guesses'.

It is questionable whether these reservations disqualify this test format. If its only purpose is to measure how well the selected target items are known, then the test may not do a very good job. But one could be interested also in more general qualitative knowledge of the lexicon. In that case it would be interesting to be able to make a difference between learners who are successful in identifying two or three associates even without knowing the stimulus word, and those who were not struck by any meaningful relationship between the nine words given in each item. Moreover, 'resourcefulness in seeking possible associates' and 'confidence to make guesses' may be seen as negative when one wants to know whether selected relationships are recognized by the learner or not. But in a more general way, such strategies seem to be helpful in normal language use and learners who exploit these means may be said to have richer vocabularies than those who do not.

Another test that tries in a somewhat similar vein to tap more qualitative aspects of lexical knowledge, the *Euralex French Tests* (Meara 1992b, henceforth EFT), does not present the kind of difficulty pointed out by Read. As there definitely is a need for more types of tests which measure depth of vocabulary knowledge and as, up to now, this particular format has gone unnoticed, I would like to present the EFT and examine to what extent they are useful. In the remainder of this paper I would like to give a detailed description of the EFT, to report on research as to their validity, their reliability and their discriminating power, and to show how they can be further developed.

TESTING HIGH LEVELS OF VOCABULARY KNOWLEDGE IN FRENCH

EFT: development and content

In the EFT, subjects are presented with a fairly large number of items which consist of two words and where the question to be answered is: are these two words related in some way or not? The items have the following format:

1 [ ]	pied: grue	2 [ ]	sarcler: bois
3 [ ]	mot: blanc	4 [ ]	joufflu: bébé

The instructions given to the testees are as follows: 'You have to decide if you can see an obvious connection between the two words'. And two examples are given, one where there is such a connection (between *vache* 'cow' and *lait* 'milk') and one where there is no relationship (between *avion* 'aeroplane' and *écrire* 'to write').

Each test consists of 60 items, 40 of which present words having some kind of relationship and 20 containing words with no relationship whatsoever (dummy items). In order to give the teacher a big choice, the complete set of EFT contains 50 tests numbered from 101 to 150. Testees should be given four tests per session. This test format was developed to assess the vocabulary knowledge of lexicographers and aimed at a very high level. The format was based on earlier EFL tests by Meara (1992a) which had been successfully used with single word items. As there was little time and money available for the construction of these tests (Meara, personal communication), it was not possible to pre-test the items or to study the validity and the reliability of this new form of test.

As can be seen in the items given as examples above, testees have to know quite infrequent items like *grue* ('crane') and *joufflu* ('chubby-cheeked'), and in that perspective the test can be taken to measure vocabulary size. But the real purpose of the test is to examine whether the test-taker knows that there is a relationship between *pied* and *grue* and between *joufflu* and *bébé,* whereas there do not exist any particular relationships between the words given in items 1 and 3. In other words, the (very) advanced learner of French L2 has to demonstrate that he knows the expression *faire le pied de grue* ('to stand about waiting'), that it is particular of a *bébé* to be *joufflu* and that no such expressions or relations exist in the two other items.

This seems indeed to be the kind of knowledge that characterizes a high level of lexical knowledge in French. As Meara (1996: 45) states, 'vocabulary size is probably the only dimension of any real importance as long as we are dealing with a small lexicon.' Beyond a threshold of some 5,000 or 6,000 words for English it is important to measure other dimensions. Indeed, to get a good score on the EFT, it does not suffice to recognize the words as belonging to the vocabulary of French (like in a yes/no-test) or to know the (or a) meaning of a word (like in most multiple choice tests). These tests are about fixed expressions, collocates and, as we will see below, about particular relationships like synonymy, antonymy, and hyponymy. Even by giving fairly frequent words like *rire* and *barbe*, which are used in the expression *rire dans sa barbe* ('to laugh to oneself'), the test taps knowledge of another type than most other vocabulary tests. This knowledge stems essentially from the domains 2 ('meaning', and especially knowledge of the semantic network) and 5 ('collocates') of the list of aspects of vocabulary knowledge given in section 1 above. In addition, it checks knowledge of lexical units that are composed of more than one content word like *soupe au lait* ('quick-tempered').

According to the *Introduction* preceding the EFT, the words used in the tests

have been selected in two ways. First, rare words were selected by working systematically through a large French dictionary and by picking only those words which were unlikely to be known by L2 speakers with a limited, i.e. non-native knowledge of French. All words that were recognized by a small group of non-native speakers were removed. In this way a list of approximately 300 infrequent words was obtained. These were given to a native speaker of French who for each of them supplied associates which were assumed to be easily recognizable for native speakers. This resulted in the first list, which consisted of 300 rare French words with an associate word for each of them.

A second list contains approximately 400 words taken from the *Français Fondamental, Premier degré* (Gougenheim *et al*. 1956). A native speaker provided associations for each of the items. These associated words had, again, to be obvious to native speakers, but not obvious to L2 speakers who knew only the basic meaning of the words. A total of 700 associated pairs were produced in this way. A third list was created by randomly pairing items of the two source lists. This list of non-associated word pairs (dummy items) was checked for fortuitous associations. Only the word pairs where no relationships were found were included in the third list.

Four types of associations were said to be taken into account. The first is of a semantic nature and includes such relationships as synonymy, hyponymy, and antonymy. The second type is based on what are called selection restrictions such as relationships between a verb and its (preferred) subject or object as in *cabrer/cheval* ('rear up/horse'). The third type aims at fixed expressions as is the case in *frotter/oreilles* (*frotter les oreilles à qn* means 'to box somebody's ears'). The last type concerns cultural aspects, an example of which is *jeu/amour* because of the title of Marivaux's play *Le jeu de l'amour et du hasard*.

It will be clear that this format is potentially very suitable to measure knowledge of a number of aspects of lexical units. Not only semantic relationships can be tested, but it is possible also to check, for instance, whether subjects recognize special relationships between verbs and their subjects or objects, or between adjectives and nouns or adverbs. Moreover, the knowledge of fixed expressions can be tested, as well as relevant cultural knowledge. Different senses of a polysemous word can very well be tested in this procedure. Items testing knowledge of compounds like *heat wave* are perfectly conceivable and do indeed appear in the EFT (see below). For English, it is not impossible for knowledge of phrasal verbs like *give up* to be tested by this type of test. Nevertheless, this format cannot test all aspects of lexical knowledge. Knowledge of morphological, syntactical, or discoursal aspects of lexical units (domains 3, 4 and 6 of the list given in the first section above) seems difficult to tap by this procedure. Finally, it is evident that only recognition, not production of lexical knowledge is measured.

In the terms of Read (1993) the EFT's constitute simple tests that produce

verifiable responses and tap data about the depth of knowledge of lexical items that are presented in isolation rather than in a normal context. Although the criteria proposed by Wesche and Paribakht (1996) were originally meant to describe tests for breadth of vocabulary knowledge, they can be applied to tests like the EFT as well. The selection of the items was done on the basis of a large dictionary, the task of the testee is to indicate whether or not there is a link between the two words of each item, and the test response format is the one just described; only the fourth dimension described by Wesche and Paribakht does not apply because there are no open-ended responses. According to Joe *et al.* (1998), the test would be classified as receptive, recognition, and rather precise.

In order to know more about the content of the tests, I have analysed six of them. Tests EFT 122–27, which, like all other tests, contained 60 items each, were randomly chosen from the set of 50 tests. This analysis revealed that some of the categories of relationships were somewhat broader than how they were announced in the *Introduction*.

First, the category 'selection restrictions' indeed contained cases like *appâter/poisson* ('to lure/fish') or *dégouliner/eau* ('to drip/water'), where verbs and their typical objects or subjects are involved. But in many cases there were no such relationships. For instance, in *obus/guerre* ('shell/war') or *gerbe/paysan* ('sheaf/farmer'), the relationships have more to do with the conceptual meaning of the words than with any restrictions in a grammatical sense. Often, one of the words of a pair has to be used or can be used when defining the other one, as is the case in *touchette/guitare* ('fret/guitar') or in *corbillard/mort* ('hearse/dead'). They seem to be better classified as free or schematic associations or as what Read (1993) calls analytic relationships. In the analyses I will treat the associations as a subcategory of the intended category 'selection restrictions'.

Secondly, the category 'fixed expressions' did not only contain words which formed expressions, like in *simple/bonjour* (*simple comme bonjour* means 'easy as pie'), but there were a considerable number of words forming compound words like *bateau/mouche* (*bateau-mouche* 'river boat') or *identité/pièce* (*pièce d'identité* 'identity paper'). Likewise, I will treat these as a subcategory of the category 'fixed expressions'.

It was not always easy to classify the items into one of the categories. A pair like *ciel/chevalier* ('heaven/knight') seems to refer to the Bible and has been classified as a cultural item. The pair *léger/air* ('light/air') has been classified as a fixed expression although the words do at the most form some kind of loose collocation. Some items turned out to be really unclassifiable and, probably, have to be considered as mistakes. One such case is *insecte/affaire* ('insect/matter' or 'insect/business'). On the other hand, one pair, where, according to the key, no relationship was anticipated, turned out to have one: *livre/jungle* ('book/jungle') as in Rudyard Kipling's *The Jungle Book*. In the vast majority of the items only the dictionary forms of the words appeared in the test. In about ten cases derived forms were given instead of canonical forms, like *assistée* instead of *assister* or *d'armes* instead of *arme*. The numbers of items pertaining

in each test to the five different categories, four types of 'positive' items and the dummy items, are given in Table 1.

Table 1. Numbers of items of different categories in the six EFT tests

	122	123	124	125	126	127	Total
Meaning relationship							56
Synonym	10	2	9	8	11	9	49
Hyponym	2	1	1	1	1	1	7
Antonym	—	—	—	—	—	—	—
Selection restrictions							81
Verbs/nouns	7	2	6	5	5	2	27
Free associations	5	18	5	10	4	12	54
Fixed expressions							88
Expressions	12	10	10	9	12	12	65
Compound words	4	3	4	4	6	2	23
Cultural aspects	—	4	2	2	1	2	11
Unclassifiable	—	—	3	1	—	—	4
Dummy items	20	20	20	20	20	20	120
Total	60	60	60	60	60	60	360

Table 1 shows that the four types of 'positive' items are unevenly represented. There does not seem to be any important reason why they should be represented more evenly, though, ideally, one might wish to find the different categories in proportions that would correspond to normal language use. The number of cultural items appears to be quite low, however, and there were no instances of antonyms in this sample. What the table does not show is that several items appeared more than once in the six tests. This was due to the fact that the tests were built up by a random selection from the data base. One item even figured in all six tests used for this study. As this was an item featuring a hyponymic relationship, this subcategory was in fact represented by only two instances. The total number of different items was 341.

Pre-testing the EFT

As a first step to study the content validity of the EFT, the six tests were given to 30 subjects having a superior command of French vocabulary. They were ten teachers of French working in Dutch universities and teacher training colleges, and twenty native speakers of French, some of them university teachers of French in France or in the Netherlands, others lexicologists

working in a French dictionary company. The idea was that if the EFT is a valid tool for the measurement of a very high level of French vocabulary knowledge, these native speakers, teachers of French and lexicologists should all obtain near perfect scores. Even if there certainly are differences between native speakers of a same language, a test of L2 vocabulary knowledge should contain items native speakers can agree upon so that there cannot be much of a discussion about the nature of the relationship between any two words of the test. So, all items should produce scores of at least 90 per cent correct answers.

The six tests were copied and were presented to these subjects on an individual basis. The results are given in Table 2. As may be seen, neither of the two groups had a perfect score or even the target minimum score of 90 per cent. In fact the native speakers obtained a score of about 83 per cent and the non-native speakers of about 80 per cent only. This is too low for groups of subjects that may be considered as experts. For that reason the content validity of this test does not seem to be satisfactory.

In order to get a clearer picture of the actual content of the tests it is useful to compare the results of the two groups. The differences between the scores of the two groups of subjects as measured by *t*-tests were not significant on the

Table 2. Results of the native and non-native groups on the six EFT tests

		Native speakers			Non-native speakers			
	N of	(*N* = 20)			(*N* = 10)			
Category	items	Mean scores	SD	%	Mean scores	SD	%	
Meaning relationship								
Synonyms	49	37.95	10.6	77	41.90	4.1	86	n.s.
Hyponyms	7	3.25	3.0	46	3.30	2.8	47	n.s.
Selection restrictions								
Verbs/nouns	27	21.20	3.9	79	22.50	2.7	83	n.s.
Free assoc.	54	42.70	8.8	79	40.10	2.9	74	n.s.
Fixed expressions								
Expressions	65	59.25	4.8	91	50.30	6.8	77	.000
Compounds	23	21.70	1.3	94	18.90	2.7	82	.001
Culture	11	9.05	1.5	82	7.30	1.9	66	.012
Unclassifiable	4	2.15	2.7	68	1.60	1.7	40	n.s.
Dummy items	120	104.35	10.6	87	102.20	11.0	85	n.s.
Total	360	300.00	27.6	83	286.40	20.1	80	n.s.

six tests taken together, although they were significantly different for three of the tests (EFT 123 at the .01 level, and EFT 124 and 127 at the .05 level). The greatest difference between the scores of the native and the non-native groups appears with the category of fixed expressions and compounds ($p < .001$). Notable differences also appear with culturally bound items ($p < .01$). The differences for all other types of items were not significant. The non-natives had a slightly better score on items where some meaning relationship is implied. This may be related to the fact that near native speakers may have less clear meaning boundaries for the words of their foreign language than native speakers and may, therefore, accept more easily associative meaning relationships like *coincer/enfermement* ('to jam/confinement'), where all Dutch subjects recognized the intended 'synonymy', whereas only half of the French subjects accepted it. This interpretation seems to be all the more acceptable because of the slightly lower score of the non-natives for the cases where they had to decide that no relationship existed between two given words.

In more than 80 cases (out of a total of 360), the two groups of subjects gave quite different answers. In cases like *cuisse/Jupiter* ('thigh/Jove'), *tesson/bouteille* ('shard/bottle'), *poule/dent* ('chicken/tooth') and *avis/population* ('notice/population') the native speakers of French almost unanimously recognized some sort of relationship whereas at most half of the non-natives answered correctly. These items could be taken to discriminate between native and non-native speakers at an extremely high level, as was the aim of the test. On the other hand, cases like *navré/chagrin* ('sorry/sorrow') and *monobloc/solide* ('cast in one piece/solid'), where a relationship was anticipated, as well as *tomber/pipe* ('to fall/pipe'), which were unrelated, were correctly answered by all the non-natives but by only half or less of the native speakers.

A number of items turned out to be very unsatisfactory. That is to say that in one or both groups half or more of the subjects gave incorrect, or at least unpredicted, answers. This was the case with 26 per cent of the items: 22 out of 56 items with meaning relationships (39.3 per cent), 24 out of 81 selection restriction items (29.6 per cent), 21 out of 88 fixed expressions and compounds (23.9 per cent), 5 out of 11 cultural items (45.5 per cent) and 21 out of 120 dummy items (17.5 per cent).

Some items seem to be too difficult, even for native speakers: only 7 out of 20 knew that a *blanchaille* ('fry') is a kind of *poisson* ('fish') or that there exists a fixed expression with *fois* ('time') and *lièvre* ('hare') (*courir deux lièvres à la fois* literally means 'chase two hares at the same time' and can be translated as 'try to keep two pots on the boil'), and only three knew the word *moitir* ('impregnate with water'), which is indeed marked as old or technical in the *Petit Robert*. In other cases, low scores were due to flaws in the tests. All French subjects as well as most of the Dutch university teachers recognized the relationship between *livre* ('book') and *jungle* ('jungle'), which had not been anticipated, and only three French subjects saw a meaning relationship between *amuser* ('to amuse') and *appât* ('bait'), which is rather vague indeed. A case like *limoger/ouvrier* ('to dismiss/worker') leads to much hesitation; the

use of this verb is severely restricted: only when high-ranking officials or equivalent are fired, is this verb used; simple workers can only be *licenciés* ('fired').

Some items appeared in more than one test. A comparison of the results of these items shows that in most cases (10 out of 16) there was no or only a very slight variability in the answers: most if not all subjects stuck to the answers they gave the first time they came across a given item. In the other cases, however, there was somewhat more instability, which was to be found among the native speakers as well as among the non-natives, but never among the two groups for one single item. For instance, *raclure/débris* ('scraping/debris') was positively answered by eight non-native teachers the first time, but by five of them only the second; *amorcer/débuter* ('begin/start') was recognized as featuring some sort of synonymy by 15 native speakers at its first appearance and by 18 the second time. This seems to stress the need for rather long tests, although it is not clear what the optimal length of this type of test would be.

In sum, only 154 of the 341 different items the six tests contained had 90 per cent scores in both groups. Approximately half of these 'good' items turned out to be dummy items, where 100 per cent scores in both groups were not rare. What this first experiment has clearly shown is that tests like the EFT have to be extensively pre-tested before they can be used with real subjects.

Reliability and discriminating power of two new tests

The second experiment aimed at studying the reliability and the discriminating power of this type of test. With the 154 items which seemed to measure, in a valid way, a high level of vocabulary knowledge, I composed two new tests of 60 items each along the same lines as were followed for the original EFT. To do so, I needed 80 positive items and 40 dummy items. For the latter category there was no problem, but for the meaning relationship, the selection restriction and the fixed expression categories, the criteria had to be somewhat lowered in a limited number of cases. This means that for some items not 18 (i.e. 90 per cent), but only17 out of 20 native speakers had given a correct answer. Table 3 gives an overview of the content of these two new tests.

These tests (see Appendix 1) were given to 30 Dutch speaking students who were at different levels of study of French at Leiden University. The students had had a minimum of six years of French at secondary level and had been studying French at university level for from one to five years. They were all native speakers of Dutch. Although not all of these students correspond to the target group the original EFT were devised for, it seemed to be interesting to investigate to what extent such tests could distinguish in a reliable way between university students at different levels of their study of French. In order to get indications about their level, the students were asked to give some

Table 3. Numbers of items of different categories in the two new tests

	A	B	Total
Meaning relationship			14
Synonyms	6	7	
Hyponyms	1	—	
Selection restrictions			26
Verbs/nouns	3	4	
Free association	10	9	
Fixed expressions			40
Expressions	16	15	
Compounds	4	5	
Dummy items	20	20	40
Total	60	60	120

supplementary information. Four aspects were taken into account: the year they began their studies, the mean of the grades they had obtained for the exams measuring language proficiency, the weekly number of hours they devoted to contacts with French and French speaking people, and the length of their stays in a French speaking country. Table 4 shows some of the results of the two tests.

Several analyses were applied to the results of these new tests. In the first place, the correlation between the two tests was at a satisfactory level (.84, $p <$ .001). The reliability of the two tests (Kronbach's alpha) turned out to be .61 for the first test and .82 for the second one. Only the latter figure can be considered to be satisfactory. As to the mean scores, which should be between 50 per cent and 90 per cent if the test is supposed to discriminate between subjects, the first test contained 33 items which did not satisfy this criterion, the second one 34. Together the two tests contained 42 items which were too easy to make a difference and 25 which were too difficult to do so. This conclusion was confirmed by the calculated item-total correlations (R_{it}-values, i.e. values of the discriminating power of each item as compared to the whole test). These values were rather low, most of them lying beneath .30 and almost none lying above .50, as would have been satisfactory. Even more disappointingly, a fair number of the R_{it}-values turned out to be negative, indicating that the better students tended to make more errors on these items and/or that weaker students tended to give the correct answers.

These data lead to the conclusion that the discriminating power of the two new tests for the subjects involved is not sufficient. It comes as no surprise, then, that no significant correlations have been measured between these tests and the level the students were at, as stated above these levels being

*Table 4. Results of the students (*N = *30) on the two new tests*

	Item type	Mean score	R_{it} value	$R_{it}2$ value		Item type	Mean score	R_{it} value	$R_{it}2$ value
Test A									
1	assoc.	.63	.05	–.01	2	comp.	.53	.11	.21
3	assoc.	.13	.40	.37	4	—	.83	–.32	
5	expr.	.36	–.27	.07	6	—	.90	–.21	
7	syn.	.46	.22	.17	8	—	.90	–.19	
9	—	.97	–.04		10	expr.	.90	.22	.19
11	—	.80	–.03		12	expr.	.87	.40	.45
13	v/n.	.17	–.02	.33	14	comp	.77	.33	.37
15	syn.	.93	–.08	–.26	16	—	.87	–.24	
17	—	.83	–.05		18	syn.	.37	.36	.35
19	—	.77	–.32		20	—	.97	–.22	
21	—	.97	–.11		22	syn.	1.00	.00	.00
23	expr.	.60	.36	.52	24	v/n.	.83	.36	.45
25	expr.	.57	.03	.10	26	syn.	.50	.46	.54
27	—	.90	–.24		28	expr.	.30	.55	.59
29	assoc.	.43	.13	.25	30	assoc.	.07	.36	.31
31	assoc.	.40		.51	32	—	.80	–.03	
33	expr.	.93	.03	.07	34	—	.87	–.03	
35	expr.	.40	.41	.60	36	assoc.	.77	.36	.46
37	hyp.	.43	.27	.44	38	assoc.	.37	.23	.32
39	—	.90	.01		40	—	.90	–.19	
41	expr.	.53	.11	.25	42	expr.	.43	–.13	.10
43	—	.83	–.16		44	expr.	.40	.49	.71
45	—	.77	–.14		46	expr.	.13	.27	.09
47	expr.	.63	.36	.41	48	comp.	.47	.41	.42
49	assoc.	.57	.43	.45	50	expr.	.47	.11	.33
51	—	.90	–.19		52	syn.	.77	.05	.16
53	expr.	.37	.29	.31	54	expr.	.67	.40	.49
55	—	.87	–.13		56	comp.	.53	.19	.19
57	assoc.	.70	.12	.21	58	—	.90	–.41	
59	v/n.	.07	.39	.50	60	assoc.	.13	.25	.45

Table 4. (cont.):

	Item type	Mean score	R_{it} value	$R_{it}2$ value		Item type	Mean score	R_{it} value	$R_{it}2$ value
Test B									
1	—	1.00	.00		2	expr.	.47	.24	.32
3	expr.	.70	.25	.29	4	—	.97	–.28	
5	syn.	.17	.64	.61	6	—	.97	–.11	
7	syn.	.70	.41	.47	8	expr.	.77	.39	.41
9	expr.	.93	.11	.10	10	assoc.	.07	.57	.58
11	—	.90	–.20		12	syn.	.07	.49	.39
13	—	.60	–.18		14	comp.	.27	.41	.41
15	expr.	.17	.40	.52	16	v/n.	.60	.43	.47
17	v/n.	.07	.57	.58	18	expr.	.23	.01	.02
19	—	.97	–.28		20	expr.	.10	.50	.49
21	assoc.	.17	.57	.59	22	syn.	.60	.68	.69
23	expr.	.93	–.03	–.01	24	assoc.	.97	.13	.17
25	—	.97	–.17		26	assoc.	.20	.61	.60
27	—	.93	–.28		28	comp.	.27	.20	.26
29	—	.93	–.22		30	assoc.	.73	.35	.42
31	v/n.	.17	.43	.40	32	expr.	.57	.36	.34
33	v/n.	.13	.41	.42	34	assoc.	.37	.63	.66
35	—	.80	–.19		36	comp.	.57	.21	.33
37	—	.73	.16		38	comp.	.83	.42	.46
39	syn.	.63	.48	.53	40	comp.	.67	.19	.17
41	expr.	.53	.09	.05	42	—	.83	.18	
43	expr.	.30	.30	.40	44	assoc.	.23	.11	.12
45	—	.87	.23		46	—	.37	.09	
47	expr.	.50	.57	.58	48	—	.80	–.38	
49	assoc.	.67	.40	.37	50	expr.	.80	.45	.49
51	syn.	.03	.56	.44	52	—	.83	.02	
53	expr.	.40	.45	.56	54	syn.	.60	.55	.60
55	—	.87	.01		56	assoc.	.33	.58	.61
57	expr.	.23	.43	.44	58	—	.83	–.23	
59	—	.87	–.25		60	—	.97	–.17	

(comp. = compound word; assoc. = free associations; expr. = fixed expression; hyp. = hyponym; syn. = synonym; v/n = selection restriction; — = no relationship)

measured as the grades they had got for their language proficiency exams, the number of years they had been studying French and the length of their stay in a francophone country. Only the second test, which had a higher reliability as measured by Kronbach's alpha, gave some significant correlations with the number of hours per week students declared to pass in reading or listening to French (resp. .51 and .53, $p < .01$).

As to the negative R_{it}-values, they almost all appear with the dummy items, i.e. the items where no relationship was anticipated (32 out of 40), whereas negative values are almost absent with the positive items (5 out of 80). In order to find an explanation for the negative R_{it}-values with 32 of the 40 dummy items, it seems sufficient to look at cases like *pelle/nez* ('shovel/nose', test A, item 58), *chaux/hameçon* ('lime/fish hook', A6) or *embrayer/désir* ('to put into gear/desire', B29) which all combine a very high mean score (90 per cent or more) with a negative R_{it}-value. In all these items, at least one word does not belong to the average vocabulary of university students of French. And when one word is unknown, there cannot be a relationship for the student; so, the answer will be 'no', which in these cases is correct. As more words will be unknown to the low scorers, these students will give more 'no' answers, which considerably lowers the discriminating power of this type of item. It should be added however that a combination of a high mean score and a negative R_{it}-value may also appear with items which contain two very frequent words like *agréable/dents* ('pleasant/teeth', A4) or *corde/facile* ('rope/easy', B4). But here again, those who are aware of fewer relationships between words, the low scorers, will tend to give a correct negative answer. The negative R_{it}-values of the five positive items are quite low, with the exception of one item, A5, which tests knowledge of the expression *avoir d'autres chats à fouetter* ('to have other fish to fry').

Looking at the satisfactory R_{it}-values (.40 and up), it becomes clear that in almost all cases less technical words are involved. Consider the following examples, presented in order of diminishing R_{it}-values: *panser/soigner* ('to put a dressing/to nurse', B22), *atout/carte* ('trump/card', B34), *agrafeuse/papier* ('stapler/paper', B26), *sillon/champ* ('furrow/field', B56), *pas/loup* ('step/wolf', B47), *obus/guerre* ('shell/war', B21), *charrue/boeuf* ('plough/ox', A28), *houille/mineur* ('coal/miner', A31). Unfortunately, only some of these items have at the same time acceptable mean scores (between 50 per cent and 90 per cent).

As was said above, many items do not satisfy the normal psychometric criteria. In fact, only 18 items combine a mean score value between 50 per cent and 90 per cent and a R_{it}-value of over .40. An item like *vitre/fenêtre* ('pane/window', A22), which has a mean score of 100 per cent and, as a consequence, a R_{it}-value of .00, is definitely too easy and should be removed from the test. The same holds for *demander/grâce* ('ask/mercy') with a mean score of 93 per cent and a R_{it}-value of –.03. On the other hand, *cabrer/cheval* ('to rear up/horse', A13) with a mean score of 13 per cent and a R_{it}-value of only .27 (or .09 after removal of the dummy items, see hereafter) seems to be

too difficult for the population and does not discriminate in a satisfactory manner either. It is less clear what the value is of an item like *tamiser/lumière* ('to filter/light', A59) which combines a mean score of 7 per cent with a R_{it}-value of .39 (but .50 after removal of the dummy items). Such items seem to be too difficult, but discriminate in a rather satisfactory way between students.

As the dummy items play a negative role as far as the discriminatory power of the test is concerned, a new analysis of the results was done with only the 40 positive items per test (those with some kind of relationship between the words of a pair). The new R_{it}-values of these 80 positive items are given in the column R_{it}-2 value in Table 4. As can be seen, there remain only three negative values, two of which are very low. This means that most of the positive items of the two tests measure the same kind of knowledge. This is indicated also by Kronbach's alpha: .84 for the first test, .90 for the second one. If the two tests are taken together, Kronbach's alpha is at .93. All these figures are satisfactory and suggest that two tests of 40 positive items each are sufficient to measure knowledge of lexical relationships of university students of French in a reliable way. It should be kept in mind, however, that such a test, as its items do not represent a real sample of French, is not necessarily a test that covers the whole lexicon of French.

For the two reduced tests taken together, there was a low but significant correlation with the mean grades the students had got for their language proficiency exams (.46, $p < .01$). Correlations with year of study or time spent in a francophone environment were not significant.

It goes without saying that not taking the dummy items into consideration creates a problem for the use of this type of test. If the subjects know how the results will be analysed, all they have to do is to put a plus in all cases. So, either one has to make sure that the subjects do not know which statistics will be applied to their results, or to think of some more sophisticated way of evaluating the results. I will come back to this issue in the next section.

The third test

In order to further analyse the usefulness of the test format inaugurated by the EFT, the best items of the two tests analysed in the preceding section were chosen to make another test; the third test. As to the positive items, no items were chosen with negative R_{it}-values, or with R_{it}-values lower than .10. All items had to have mean scores between 50 per cent and 90 per cent, or, in case of R_{it}-values above .20, between 40 per cent and 90 per cent. The dummy items had to have a mean score between 50 per cent and 90 per cent and R_{it}-values of –.15 or above. This leads to a test of 70 items, 58 positive ones and 12 negative ones (see Appendix 2).

This test was given to 29 students who were at different stages of their study of French at the same Dutch University. Table 5 presents the results of the third test. As can be seen, 19 of the 58 positive items have satisfactory mean

scores (between 50 per cent and 90 per cent) and another 10 have acceptable mean scores (between 40 per cent and 50 per cent). There are no mean scores higher than 90 per cent, which means that none of the items must be considered to be too easy for the subjects involved. Half of the positive items have mean scores which lie below 40 per cent. As to the 12 dummy items, 6 have satisfactory mean scores (between 50 per cent and 90 per cent) and 6 have mean scores of more than 90 per cent, two of them giving a 100 per cent score. The overall mean of the test is .50. This means that the test as a whole is rather difficult for the subjects.

Table 5 also shows that 24 out of 58 positive items have R_{it}-values of at least .40, whereas another 14 are superior to .30. Four positive items have negative R_{it}-values, three of them being near to zero and only one (item 66) behaving in a rather unexpected way: *sourd/pot* ('deaf/pot': as deaf as a post) had been one of the better items in the second test. Four of the 12 dummy items have negative R_{it}-values, two of which (items 33 and 54) seem to be of some importance (–.22 and –.31 respectively). The overall reliability of the test as measured by Kronbach's alpha is .84, which means that the test measures vocabulary knowledge in a consistent way.

As was the case with the two tests which were discussed above in the section entitled 'Reliability and discriminating power of the two new tests', the results of the third test were correlated with other data which had been provided by the subjects, such as year of study, grades for translation, grammar, speaking and listening and contacts with French and the French. There were significant correlations with 'year of study' and 'grammar' (.44 and .46 resp., $p < .05$); there was a highly significant correlation with speaking skill (.61, $p < .01$). The correlation with translation French–Dutch was at .37 ($p = .09$). Correlations with translation Dutch–French and listening or with time spent in a francophone environment were not significant. Although most of the correlations are not high, as a whole, the third test can be taken to be a better predictor of language proficiency than the two tests of the previous section.

In order to examine the role of the dummy items, some additional statistics were computed. In the first place, results were calculated of the positive items only, leaving out all dummy items. This lowered all correlations with the language proficiency data, year of study and contact with French with one to five points. This means that, in spite of the negative R_{it}-values of a number of dummy items, on the whole this type of item adds to the discriminating power of the test. On the contrary, subtracting points for all positively (i.e. erroneously) answered dummy items had a positive effect on the level of correlations. After having compared several weights for the dummy items, counting each positively answered dummy item as two errors gave the best results with a highly significant correlation of .64 with speaking skill. This suggests a solution for the scoring problem. If test takers know that all positively answered dummy items will be counted as two errors, they will have to choose carefully the ones they consider to have no relationship. As to

*Table 5. Results of the students (*N = *29) on the third test*

	Item type	Mean score	R_{it} value		Item type	Mean score	R_{it} value
1	comp.	.45	.33	36	assoc.	.17	.35
2	assoc.	.07	.62	37	expr.	.48	.36
3	syn.	.24	.33	38	syn.	.31	.48
4	syn.	.41	.11	39	syn	.55	.32
5	—	.72	.07	40	expr.	.48	.33
6	expr.	.83	–.06	41	comp.	.24	.41
7	v/n	.07	.60	42	assoc.	.35	.42
8	comp.	.83	.01	43	expr.	.10	.42
9	—	1.00	.00	44	v/n	.34	.31
10	syn	.28	.58	45	assoc.	.31	.4
11	expr.	.66	.27	46	syn.	.52	.25
12	v/n	.76	.26	47	comp.	.76	.44
13	expr.	.52	.48	48	assoc.	.72	.29
14	assoc.	.21	.27	49	v/n	.24	–.08
15	assoc.	.24	.31	50	expr.	.48	.48
16	—	.93	.02	51	v/n	.07	.17
17	—	.90	.00	52	assoc.	.28	.34
18	—	.90	.00	53	comp.	.59	.16
19	expr.	.38	.35	54	—	.93	–.31
20	assoc.	.66	.44	55	—	.86	–.05
21	hyp.	.38	.28	56	comp.	.76	.27
22	assoc.	.14	.61	57	syn.	.72	.37
23	expr.	.48	.40	58	comp.	.52	–.05
24	expr.	.48	.21	59	expr.	.10	.42
25	expr.	.62	.32	60	—	.97	–.25
26	comp.	.34	.44	61	—	.52	.10
27	assoc.	.66	.46	62	expr.	.28	.27
28	expr.	.21	.55	63	assoc.	.55	.44
29	syn.	.59	.44	64	expr.	.38	.50
30	expr.	.52	.27	65	—	1.00	.00
31	expr.	.17	.43	66	expr.	.10	–.29
32	expr.	.45	.34	67	expr.	.34	.22
33	—	.90	–.22	68	syn.	.45	.46
34	comp.	.72	.33	69	—	.93	.34
35	assoc.	.31	.61	70	assoc.	.24	.45

(comp. = compound word; assoc. = free association; expr. = fixed expression; hyp. = hyponym; syn. = synonym; v/n = selection restriction; — = no relationship)

test construction, it seems advisable to have a proportion of one third of dummy items in a test, as was the case in the original EFT.

DISCUSSION AND CONCLUSION

Up to now, no entirely satisfactory tests of vocabulary knowledge have been devised. Read (1988: 17) comes to the conclusion that checklists, multiple choice tests as well as tests where words are to be matched with their definitions are all formats which can be 'criticized as inadequate indicators of whether the word is really known', and Meara (1996: 38) expresses the same opinion. It should be remembered that both authors are speaking about measures of vocabulary size, the dimension that seems to be the easier one when it comes to measuring lexical knowledge. In the meantime it is recognized that there are other aspects that should be paid attention to and that there is a need for 'vocabulary measures of differing sensitivity, so that each word is tested several times in different ways' (Joe *et al.* 1998; see also Wesche and Paribakht 1996).

Meara (1996: 49) suggests that 'it might be possible to develop an alternative way of characterising a lexicon, one that is largely, though not entirely, independent of size', i.e. in terms of organization. The most concrete attempt at measuring quality rather than size, or depth rather than breadth, is Read (1993). But at the end of his paper, Read declares that he is not entirely certain about the extent to which 'this format has achieved the goal of measuring depth of knowledge of particular words' (see also Read 1997: 317).

One of the major problems in the construction of vocabulary tests is the absence of any standard. In the first place there is a 'lack of agreement among theoreticians, researchers and questionnaire respondents about what it means to 'know' a word', as Wesche and Paribakht (1996: 14) put it. And they go on (p. 32) by saying that there is, on the one hand, a 'lack of theoretical consensus about the nature and course of development of L2 vocabulary knowledge' and, on the other hand a 'lack of alternative instruments that aim to measure the level of knowledge of individual words [which] makes it difficult to estimate concurrent validity'. So, on the theoretical level as well as on the practical level, we are confronted with an empty space as far as vocabulary acquisition is concerned. This situation could easily turn into a vicious circle where everyone is waiting for the others: those who want to define vocabulary knowledge want to be able to measure it, but at the same time test constructors will only be able to develop valid and reliable tools if it is clear what has to be understood by vocabulary knowledge.

It is in this context that every new attempt at measuring aspects of lexical knowledge must be welcomed and taken seriously. That is what has been done in this paper. And it will be clear that nothing very definitive can be said about the EFT or the ideas they were built on. However, what seems clear is that the EFT tests in their original state are not sufficiently valid: subjects who are known to have the level of vocabulary knowledge required to get high

scores on these tests do not obtain (near) perfect scores. However, two types of items, those which measure knowledge of fixed expressions and those where some kind of cultural knowledge is involved, do discriminate between native speakers and very high proficiency non-native speakers like university teachers of French.

Two tests developed with the best items of six of the EFT tests did not turn out to be sufficiently reliable tools to measure the vocabulary knowledge of university students. The overall reliability of only one of the tests was satisfactory. A further analysis of the tests revealed that the R_{it}-values of the dummy items were negative in the vast majority of the cases, meaning that these items measured something other than what was being measured by the rest of the test. If the dummy items are removed, more homogeneous tests result, which give better overall reliabilities (Kronbach's alpha at .84 and .90 respectively). When the two tests without dummy items are taken together, they give a highly reliable image of vocabulary knowledge (Kronbach's alpha at .93) of university students of French. They have, in addition, a modest but significant correlation with grades obtained for proficiency in the language.

Concerning the latter point, it is difficult to say what the desirable level of the correlations between a test of qualitative lexical knowledge and marks for grammar or translations would be. One of the problems is that we do not know to what degree general language proficiency is determined by vocabulary knowledge. So, for the moment, positive correlations with speaking, listening, reading or writing may be taken to be in favour of the content of a vocabulary test. If one or more of these correlations are (highly) significant, this could shed light on the (type of) vocabulary knowledge measured as well as on the special link with the particular type of proficiency.

However, the tests that have been discussed in this paper do not seem to be useful as practical tools for measuring levels of vocabulary knowledge in an educational context when the dummy items are taken out or when these items are not taken into account in the analysis of the results. Moreover, ignoring the dummy items means that potentially useful information about the lexical knowledge of the subjects is not considered. All this means that it is important to maintain the format of the tests and to find a solution for the statistical treatment of the dummy items.

In the light of this argument, a third test was composed with the best items of the two new tests discussed above. As the reliability of the test as a whole, i.e. including the dummy items, was at a satisfactory level, it can be said that this test measures in a consistent way some aspects of depth of lexical knowledge on a fairly high level. As counting positively answered dummy items as two errors had a positive effect on the correlations with measures for language proficiency and contact with the language, this method could be used in educational contexts. This method hints at a proportion of one third of dummy items in a test.

This paper has tried to demonstrate that an EFT-type of test is potentially a valid, reliable and useful tool for the measurement of specific aspects of lexical

knowledge in an L2, especially at rather high levels. It has shown that before using this type of test it is important to ascertain its validity by pre-testing the items with native speakers of the language and/or with other groups of speakers who definitely have the level of knowledge that is going to be measured. Next, it has been pointed out that it is necessary to study the reliability of the tests by taking into account psychometric data about the items. With the help of these data it is possible to compose tests which measure the relevant knowledge in a reliable way and which are able to discriminate between groups of students. By analysing the results obtained and by improving step by step its content, this type of test, which was at first meant for a restricted set of non-native speakers of French with a very high proficiency level, seems to be fairly adaptable to different testing environments. As choices can be made concerning the aspects of vocabulary knowledge that one wants to test, it seems worthwhile to carry on research on this format.

Final version received January 2000

ACKNOWLEDGEMENTS

I would like to thank Dr Lydius Nienhuis (University of Utrecht) and Dr Vincent van Heuven (Leiden University) for their help with the statistics.

APPENDIX 1—THE TWO NEW TESTS

Test A

1. isoloir: élection	[ ]	2. bateau: mouche	[ ]
3. affût: chasseur	[ ]	4. agréable: dents	[ ]
5. chat: fouetter	[ ]	6. chaux: hameçon	[ ]
7. barbe: ennui	[ ]	8. potelé: corde	[ ]
9. aviron: singe	[ ]	10. dîner: chandelle	[ ]
11. bâtisse: bébé	[ ]	12. bouillir: colère	[ ]
13. cabrer: cheval	[ ]	14. champ: honneur	[ ]
15. décorer: médaille	[ ]	16. délabré: langue	[ ]
17. allumer: petit	[ ]	18. amorcer: débuter	[ ]
19. amuser: grâce	[ ]	20. enluminure: manger	[ ]
21. flétrir: multiplier	[ ]	22. vitre: fenêtre	[ ]
23. vue: nez	[ ]	24. former: jeunesse	[ ]
25. frotter: oreilles	[ ]	26. gâchis: gaspiller	[ ]
27. champ: clou	[ ]	28. charrue: boeuf	[ ]
29. gerbe: paysan	[ ]	30. hâle: soleil	[ ]
31. houille: mineur	[ ]	32. idée: pièce	[ ]
33. jour: ouvrable	[ ]	34. ciel: bec	[ ]
35. coq: gaulois	[ ]	36. lézarde: mur	[ ]

37. babouin: singe [ ]
38. bail: appartement [ ]
39. manger: court [ ]
40. mentir: poule [ ]
41. monde: fou [ ]
42. nez: moutarde [ ]
43. octroi: sacré [ ]
44. oreilles: casser [ ]
45. parent: tabac [ ]
46. délier: langue [ ]
47. doigt: oeil [ ]
48. eau: javel [ ]
49. embolie: sang [ ]
50. peau: neuve [ ]
51. plein: corde [ ]
52. décorer: honneur [ ]
53. rire: barbe [ ]
54. simple: bonjour [ ]
55. soûl: hypocrite [ ]
56. soupe: lait [ ]
57. pelage: animal [ ]
58. pelle: nez [ ]
59. tamiser: lumière [ ]
60. vanne: robinet [ ]

Test B

1. bifteck: mer [ ]
2. blanc: yeux [ ]
3. colère: piquer [ ]
4. corde: facile [ ]
5. crinière: chevelure [ ]
6. jaunisse: panier [ ]
7. journal: canard [ ]
8. matin: petit [ ]
9. médaille: revers [ ]
10. mité: vêtement [ ]
11. boudin: insister [ ]
12. chanson: rengaine [ ]
13. chien: chevalier [ ]
14. civet: lapin [ ]
15. jaune: rire [ ]
16. croquer: pomme [ ]
17. dégouliner: eau [ ]
18. passer: tabac [ ]
19. nord: casser [ ]
20. nord: perdre [ ]
21. obus: guerre [ ]
22. panser: soigner [ ]
23. demander: grâce [ ]
24. démarrage: voiture [ ]
25. abside: bois [ ]
26. agrafeuse: papier [ ]
27. fondre: poignet [ ]
28. fossé: générations [ ]
29. embrayer: désir [ ]
30. éplucheur: pommes de terre [ ]
31. fêler: crâne [ ]
32. fer: croiser [ ]
33. appâter: poisson [ ]
34. atout: carte [ ]
35. fer: chien [ ]
36. feu: joie [ ]
37. blouse: argent [ ]
38. boîte: nuit [ ]
39. fouiller: recherches [ ]
40. identité: pièce [ ]
41. induire: erreur [ ]
42. jambe: amour [ ]
43. mot: mâcher [ ]
44. moulinet: pêche [ ]
45. mouton: âge [ ]
46. navré: lac [ ]
47. pas: loup [ ]
48. escamotage: mer [ ]
49. dévisser: tourne-vis [ ]
50. doux: mot [ ]
51. pitre: amusant [ ]
52. poule: image [ ]
53. puce: oreille [ ]
54. rétrécir: diminuer [ ]
55. sauter: drapeau [ ]
56. sillon: champ [ ]
57. sourd: pot [ ]
58. tambour: voile [ ]
59. peur: pouce [ ]
60. piètre: neige [ ]

APPENDIX 2—THE THIRD TEST

1. bateau: mouche	[ ]	2. affût: chasseur	[ ]
3. barbe: ennui	[ ]	4. gâchis: gaspiller	[ ]
5. bâtisse: bébé	[ ]	6. bouillir: colère	[ ]
7. cabrer: cheval	[ ]	8. champ: honneur	[ ]
9. allumer: petit	[ ]	10. amorcer: débuter	[ ]
11. vue: nez	[ ]	12. former: jeunesse	[ ]
13. charrue: boeuf	[ ]	14. gerbe: paysan	[ ]
15. houille: mineur	[ ]	16. idée: pièce	[ ]
17. ciel: bec	[ ]	18. manger: court	[ ]
19. coq: gaulois	[ ]	20. lézarde: mur	[ ]
21. babouin: singe	[ ]	22. bail: appartement	[ ]
23. monde: fou	[ ]	24. oreilles: casser	[ ]
25. doigt: oeil	[ ]	26. eau: javel	[ ]
27. embolie: sang	[ ]	28. peau: neuve	[ ]
29. décorer: honneur	[ ]	30. blanc: yeux	[ ]
31. rire: barbe	[ ]	32. simple: bonjour	[ ]
33. soûl: hypocrite	[ ]	34. soupe: lait	[ ]
35. pelage: animal	[ ]	36. vanne: robinet	[ ]
37. colère: piquer	[ ]	38. crinière: chevelure	[ ]
39. journal: canard	[ ]	40. matin: petit	[ ]
41. civet: lapin	[ ]	42. agrafeuse: papier	[ ]
43. jaune: rire	[ ]	44. croquer: pomme	[ ]
45. obus: guerre	[ ]	46. panser: soigner	[ ]
47. fossé: génération	[ ]	48. éplucheur: pomme de terre	[ ]
49. fêler: crâne	[ ]	50. fer: croiser	[ ]
51. appâter: poisson	[ ]	52. atout: carte	[ ]
53. feu: joie	[ ]	54. jambe: amour	[ ]
55. blouse: argent	[ ]	56. boîte: nuit	[ ]
57. fouiller: recherches	[ ]	58. identité: pièce	[ ]
59. mot: mâcher	[ ]	60. mouton: âge	[ ]
61. navré: lac	[ ]	62. pas: loup	[ ]
63. dévisser: tourne-vis	[ ]	64. doux: mot	[ ]
65. poule: image	[ ]	66. sourd: pot	[ ]
67. puce: oreille	[ ]	68. rétrécir: diminuer	[ ]
69. sauter: drapeau	[ ]	70. sillon: champ	[ ]

REFERENCES

Bogaards, P. 1994. *Le vocabulaire dans l'apprentissage des langues étrangères*. Paris: Crédif–Hatier/Didier (Coll. LAL).

Bogaards, P. 1996. 'Dictionaries for learners of English.' *International Journal of Lexicography* 9: 277–320.

Bogaards, P. (forthcoming). 'Lexical units and the learning of foreign language vocabulary.' *Studies in Second Language Acquisition*.

Chapelle, C. A. 1994. 'Are C-tests valid measures for L2 vocabulary research?.' *Second Language Research* 10: 157–87.

Cronbach, L. J. 1942. 'An analysis of techniques for diagnostic vocabulary testing.' *Journal of Educational Research* 36: 206–17.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Gougenheim, G., P. Rivenc, R. Michéa, and **A. Sauvageot.** 1956. *L'élaboration du français fondamental (1er degré)*. Paris: Didier.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Joe, A., P. Nation, and **J. Newton.** 1998. 'Vocabulary test difficulty and item types' unpublished paper.

Laufer, B. and **P. Nation.** 1995. 'Vocabulary size and use: Lexical richness in L2 written production.' *Applied Linguistics* 16: 307–22.

Laufer, B. and **P. Nation.** 1999. 'A vocabulary-size test of controlled productive ability'. *Language Testing* 16: 33–51.

Meara, P. 1992a. *EFL Vocabulary Tests*. Swansea: Centre for Applied Language Studies.

Meara, P. 1992b. *Euralex French Tests*. Swansea: Centre for Applied Language Studies.

Meara, P. 1996. 'The dimensions of lexical competence' in G. Brown, K. Malmkjaer, and J. Williams (eds.): *Performance and Competence in Second Language Acquisition*. Cambridge: Cambridge University Press. 35–53.

Meara, P. and **B. Buxton.** 1987. 'An alternative to multiple choice vocabulary tests.' *Language Testing* 4: 142–54.

Nation, P. 1983. 'Testing and teaching vocabulary.' *Guidelines* 5: 12–25.

Nation, P. 1990. *Teaching and Learning Vocabulary*. New York: Newbury House.

Read, J. 1988. 'Measuring the vocabulary knowledge of second language learners.' *RELC Journal* 19/2: 13–26.

Read, J. 1993. 'The development of a new measure of L2 vocabulary knowledge.' *Language Testing* 10: 355–71.

Read, J. 1997. 'Vocabulary and testing' in N. Schmitt and M. McCarthy (eds.): *Vocabulary. Description, Acquisition and Pedagogy*, Cambridge: Cambridge University Press. 303–20.

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

Singleton, D. and **D. Little.** 1991. 'The second language lexicon: some evidence from university-level learners of French and German.' *Second Language Research* 7: 62–81.

Wesche, M. and **T. S. Paribakht.** 1996. 'Assessing second language vocabulary knowledge: depth versus breadth.' *The Canadian Modern Language Review* 53: 13–40.