Using Category Generation Tasks to Estimate Productive Vocabulary Size in a Foreign Language

Shadan Roghani
Swansea University, UK

James Milton*
Swansea University, UK

Abstract

This paper reports an investigation into whether a test of productive vocabulary size using a category generation task can be useful and effective. A category generation task is a simple task where learners are asked to name as many as words as they can from a prescribed category such as *animals* or *body parts*. The virtue of this approach is that it potentially allows an estimate of productive vocabulary size, comparable to receptive size estimates, to be made. Four such tasks were trialled on 92 learners ranging from elementary to advanced level. Subjects also took Nation's Productive Vocabulary Levels Test (PVLT) (2001) and Meara & Milton's X-Lex (2003). The results suggest that category generation tasks can produce vocabulary size estimates and these are comparable in size with PVLT and about one third of the size of a receptive vocabulary size estimate (X-Lex). The tests appeared very reliable and can distinguish between learners of different levels of performance. There are still issues to be resolved concerning the tasks which can be used and the volumes of vocabulary they can potentially obtain. Factor analysis suggests the receptive and all the productive tasks test a single factor.

Key words: productive vocabulary, vocabulary size, category generation task, vocabulary assessment, frequency vocabulary bands

Introduction

The acquisition of vocabulary knowledge, that is growing a lexicon of an appropriate size and quality, is crucial to language learning success. Since it is an aspect of language knowledge which is so important, it would make sense to measure and monitor its development among learners, and where this is done it appears that measurements of knowledge can be very useful. So, for example, estimates of vocabulary size correlate well with performance in all the language skills and in formal exams (e.g. Stæhr, 2008; Milton et al, 2010). Learners with larger vocabularies tend to perform better than those with smaller vocabulary knowledge in these activities. Approximate vocabulary sizes have been identified as requirements for passing formal exams such as Cambridge FCE in English, and have been linked to hierarchies of communicative levels as in the CEFR (Milton, 2010; Milton & Alexiou, 2009). Because vocabulary is so important perhaps it is not surprising that students identify shortcomings in their L2 vocabulary knowledge as a principle obstacle to comprehension (Laufer, 1989). The importance of vocabulary is such that Long & Richards (2007, p.xii) suggest that 'vocabulary can be viewed as the core component of all the language skills.'

**Tel: (44) 1792 295678; E-mail: j.l.milton@swansea.ac.uk; Department of Applied Linguistics, Swansea University, Singleton Park, Swansea SA2 8PP, UK.*

While clearly insightful, there is a feeling among academics that making this kind of measurement can be a complicated business. Vocabulary knowledge, it seems, is multifaceted. It can include knowledge of both the written and oral forms of words. It includes possessing a link between a word form and its meaning including the associations which a word can carry and which can vary from one language to another. It can include a knowledge of how words can combine into collocations and idioms, and a knowledge too of when, and when not, to use some of these words. Vocabulary researchers usually make a distinction between vocabulary size or breadth, the number of words a learner knows, and vocabulary depth, how well these words are known and how well and idiomatically they can be used. It can also include making a distinction between receptive and productive knowledge, an observation that goes back at least as far as Palmer (1921). Palmer identified a difference between the words learners can recognise, what is called today receptive or passive vocabulary, and the words a learner can use and communicate with, a sub-set of the receptively known words which learners can readily call to mind for use in speech and writing and referred to today as productive or active vocabulary. He suggested these different types of word knowledge should be assessed separately. Different tasks, it seems, appear to activate different kinds of vocabulary knowledge (Webb, 2005), and different kinds of vocabulary knowledge can impact on the different language skills. For example, Milton & Riordan (2006) observed that knowledge of words in their oral form can be measured separately from word knowledge in written form and that oral word recognition predicts success in speaking tests while the ability to recognise words in their written form does not.

These different dimensions and features of vocabulary knowledge cannot, of course, be entirely unrelated. Possession of a large receptive vocabulary is a precondition of having a large productive vocabulary, for example, and the various dimensions of knowledge generally correlate quite well with each other as is noted by Fitzpatrick & Milton (2014). There are even arguments that suggest they can be collapsed into a single dimension of vocabulary knowledge. Vermeer (2001) argues that breadth and depth are essentially the same construct. Meara (1997) argues that automaticity in word production is a product of the number of links between words, a product of depth therefore. Fitzpatrick & Milton (2014, p.177) in considering the strength of the inter-relationship between the elements of vocabulary knowledge speculate that it may be possible, 'through frequency (Ellis, 2002a; 2002b) to explain the driver behind all the aspects of knowledge in [Nation's] table.' Nonetheless, multiple testing of vocabulary knowledge is often advocated so that a learner's knowledge can be more fully characterised (e.g. Nation, 2007; Richards & Malvern, 2007). While there seems general agreement that using multiple tests is desirable it is not clear that this is actually done outside the realm of specialist researchers. Perhaps this is because the standard tests of vocabulary are relatively few and are limited, largely, to testing receptive vocabulary breadth. This paper is particularly concerned with assessing the potential for a test which measures productive knowledge; how many words do learners have that they can easily activate and use for communication, in the hope that this will make the process of multiple testing more practical.

There are several well recognised tests is the area of receptive vocabulary size, but well-established tests are lacking in other areas of vocabulary knowledge such as productive vocabulary knowledge. Receptive vocabulary size, or breadth, testing attempts to estimate how many words in the foreign language a learner can recognise, and this type of testing is usually distinguished from vocabulary depth testing which attempts to assess how well these words are known and whether they can be used appropriately Receptive breadth tests have the advantage in their creation that the writer can control the items being tested and make a principled selection of words from which a good estimate of knowledge can be made. Both Nation's Vocabulary Size Test (VST) (2012) and Meara & Milton's X-Lex (2003) work in this way and sample words across the frequency bands and this is used to form an estimate of vocabulary size. These tests also have the advantage that they do not have to be customised to the first language of the learners and can be quick to deliver and are easy to mark. Nation's VST uses a multiple choice format where the learners select a meaning for a test word from a choice of four explanations and where the explanations are 'in much easier language than the tested word' (Nation, 2012, p.3). The checklist format in Meara & Milton's X-Lex is particularly minimalist requiring only that the testee identifies words that they recognise in a list, and the computer version of this takes only a few minutes to deliver and marks itself. With

both tests it appears relatively straightforward to produce parallel forms of the tests and the different forms are reported to be equivalent (Nation, 2012; David, 2008).

However, even these tests have their drawbacks. Nation (2012) reports that VST may under-estimate where learners are not motivated to perform on the test, but this could be said of any form of assessment. A more serious consideration is the potential for the test to over-estimate where learners are prepared to use guesswork to provide answers to words they do not know. The multiple choice format means that there is a one in four chance of getting the right answer by guesswork and there appears to be no mechanism for recognising where this is occurring and adjusting for it when it does occur. X-Lex does have such a mechanism and includes false words, and, where the testee identifies these as known words, an arithmetic formula is applied and the score is reduced. But X-Lex's simple checklist method is also prey to potential problems especially in terms of dealing with learners' uncertainty over their knowledge of a word. This form of test takes no account of partial or incomplete knowledge, and low level learners in particular are often unsure over things like spelling and may not, therefore, be able to represent the knowledge that they have. Nonetheless, both tests are reported to be robust and reliable. In an ideal world a test of productive vocabulary knowledge would have the good qualities of the receptive tests and would be easy to use and capable of accessing a sufficient and principled sample of the learner's vocabulary from which to form a good estimate of size. Ideally it should be able to demonstrate good reliability so test and retest scores, for example, should not differ significantly if there is no change in the vocabulary knowledge being tested. It should be able to demonstrate the same kinds of construct validity that receptive tests have, as in the ability to draw on a principled sample of words from across the frequency bands so that a good estimate of size can be made. It should possess good concurrent validity and correlate appropriately with other scores of the same or similar quality. So, a good productive test, if it is working well, should correlate with other tests of productive vocabulary size and should probably correlate too, although perhaps less well, with receptive vocabulary size which is generally considered a different although related construct.

Well recognised productive tests are harder to find than receptive tests. This may be because in many productive tasks, the choice of words is that of the testee and this may prevent a useful sample of words being created from which meaningful conclusions about vocabulary size or knowledge can be drawn. Thus, measures of lexical diversity and sophistication (e.g. Meara & Bell, 2001, P-Lex) appear sensitive to genre (van Hout & Vermeer, 2007) so the scores they produce may say more about the nature of the text rather than the lexicon which produced it. These measures are also sensitive to length and a minimal length, usually several hundred words, is needed before stable results are achieved (e.g. Meara & Bell, 2001). These approaches do not generally produce an estimate of size but the exception to this is Meara & Miralpeix's V-Size (2008) which analyses a testee's text and calculates the proportions of vocabulary occurring in five frequency bands to produce a curve. This curve can then be compared with curves from other texts where the size of the writer's lexicon is known and an estimate of the testee's lexical size can be made. Meara & Miralpeix's initial conclusions are that this approach is not sensitive to genre or to the length of text and that it can discriminate between learners of different ability levels. The idea is an interesting one but our experience is that the scores it produces are rather erratic and more work is probably needed to demonstrate the reliability of this approach.

Other approaches to productive vocabulary testing use controlled methods for eliciting knowledge. Laufer & Nation's Productive Vocabulary Levels Test (PVLT) (1999) takes a sample of words from the second, third, fifth and tenth 1000 word frequency ranges, and from the university word list as the target vocabulary for their test. Students are presented with a sentence giving context with the target word missing from the context, although the initial letters of the target are provided. Testees fill in the missing word. This approach has the considerable merit that its sample of words is directly equivalent to Nation's receptive Vocabulary Levels Test (Nation, 2001) and so productive and receptive scores ought to be directly comparable. The approach has been criticised, however, in that the degree of contextualisation may be so great that it becomes a receptive test in another form (Webb, 2008). This strikes at the heart of the issue in the creation of a test of productive vocabulary knowledge. Productive performance requires some kind of prompt and there is no agreed construct of productive knowledge to guide us as to how rich or minimal in contextualisation such a prompt should be. Webb (2008) considers a less

rich context in testing, therefore and suggests the merit of a translation test where the testees are presented with a prompt in their native language to elicit a translation into the foreign language target word. The approach is a simple one which ought to allow the test writer to make the kind of sample of knowledge that an estimate of vocabulary size could be drawn from. In terms of practicality, however, this approach will not be so straightforward in, say, a class of learners from many different first language backgrounds and where multiple different forms of the test will be needed. It seems there is still the opportunity for a convincing methodology to emerge in this area to produce meaningful and useful estimates of productive vocabulary size which, like the receptive tests described above, are simple enough to be used by learners from all language backgrounds and with a simple enough prompt to avoid replicating a receptive test in another form.

The research presented in this paper aims to access and measure productive vocabulary size using a new test format to see if category generation tasks can be a useful addition to testing in this area.

What Are Category Generation Tasks?

A category generation task is a simple task where the student is asked to name as many as words as they can from a prescribed category such as animals, body parts, clothes or furniture. This approach to word elicitation is widely used in psychology research and produces reliable scores which can be used to provide evidence of, for example, cognitive development or language impairment (Izura, Hernández-Muňos & Ellis, 2005). While this approach has been used with bilingual children (McKinney, 2009), it does not appear to have been used among second and foreign language learners to produce estimates of vocabulary size.

In the context of foreign language learning this approach does raise issues as to whether testing knowledge of lexical sets in this way can provide a good estimate since these are staple thematic areas generally addressed in elementary learning materials. Where knowledge of these areas is specifically taught, it may not accurately reflect knowledge of vocabulary overall. It might be argued too that a testing approach based on lexical sets might have an unwanted backwash in encouraging the teaching of vocabulary through lexical sets, a technique currently thought to be less than optimal (e.g. Tinkham, 1997). However, teaching materials, if they are to be coherent and usable, must have some thematic organisation and a testing approach that reflects this might be thought desirable. It should be noted that the research evidence with suggests that teaching vocabulary in semantically unrelated sets promotes better retention, always shows too that teaching through lexical sets is effective. We would argue, also, that this is a legitimate productive task since lists are widely used by all language users for example for shopping or when packing for holidays and is therefore a meaningful way of accessing productive knowledge. It is a task which requires minimal explanation and is equally applicable to learners regardless of their language background. We are aware that, notwithstanding potential shortcomings, it is a technique currently used in EFL where the vocabulary knowledge of very young and low level learners is tested and where more complex production is impractical. It is part of the purpose of the research presented here to assess whether these issues prevent the technique from producing good estimates of productive vocabulary size.

The category generation task format potentially offers the chance to gain an estimate of productive vocabulary size comparable to receptive size measures. Language learners have a tendency to learn frequently occurring vocabulary before less frequent items (Milton, 2007) and this provides a rationale for receptive vocabulary tests where the selection of items focuses on the initial frequency bands. In respect of the category generation tasks, frequency lists as used in Cobb's website (Cobb, 2014) can provide us with items from each category divided by frequency band. So, for example, the BNC/COCA Cobb uses lists include six animals in the first 1000 word band. If, in producing a list of animals the testee names all these six animals then for the purposes of estimating vocabulary size it might be assumed that all the words in this 1000 word band are known. If only three are produced then it might be estimated that only 500 of this 1000 word band are known. By examining knowledge of the words from each category which occur in the more frequent ranges a workable estimate of overall size can be made.

Research Questions

The intention in this study is to use four category generation tasks with EFL learners and to use the words that testees produce to calculate estimates of productive vocabulary size which might be seen as equivalent to the receptive vocabulary size estimates produced by X-Lex. The broad aim, therefore, is to examine whether these estimates can be fairly described as believable, reliable and valid. Do category generation tasks have potential as useful measures of vocabulary knowledge?

To achieve this broad aim we have set a number of specific research questions.

1. Is there a frequency effect in learning to suggest that a test targeted on the first five 1000 words bands is appropriate in a productive test?
2. Does the test produce sufficient data for estimates of size to be made?
3. Do the scores from parallel forms of the test suggest that the test is reliable? Do they produce estimates which are similar in size and which correlate?
4. Are the scores comparable with other equivalent tests of vocabulary size and knowledge: Laufer & Nation's PVLT (1999) and Meara & Milton's X-Lex (2003)?
5. Are estimates on the test capable of distinguishing between learners at different levels of knowledge and performance: beginner, intermediate and advanced levels?
6. Do these tests and PVLT access a single factor of knowledge, productive vocabulary size, and can this be distinguished from a receptive vocabulary size measure X-Lex?

Method

Participants

A total of 92 EFL learners were tested in a foreign language teaching institute in Iran. The learners came from three difference levels of knowledge: basic, intermediate and advanced levels as categorised by the institute. The 92 learners comprised 43 male and 49 female participants, were aged between 15 and 40, and were distributed among the three levels as shown in Table 1.

Table 1
Participant Levels

Level	Basic	Intermediate	Advanced	Total
Number	36	23	33	92

The Tests

Four category generation tasks were used: *animals*, *clothes*, *body parts* and *furniture*. These categories are described by Izura et al (2005, p.386) as 'commonly used in cognitive, neuropsychological and linguistic research' and which proved capable of prompting considerable language output from the participants.

Laufer & Nation's Productive Levels Test version C (Nation, 2001, p.425-428) was used as a second test of productive vocabulary knowledge. Scores from versions of Nation's VLT are widely used as a proxy for vocabulary size (e.g. Stæhr, 2008). The entire test was not administered and only the 2,000, 3,000 and 5,000 levels were used. This was converted to a productive vocabulary size estimate out of 5,000 using the formula:

$$\frac{\text{2000 level score} * 2000}{18} + \frac{\text{3000 level score} * 1000}{18} + \frac{\text{5000 level score} * 2000}{18} = \text{size}$$

A paper version of Meara & Milton's X-Lex (2003) was used as a second measure of vocabulary knowledge. This version tests 20 words in each sample across each of the five most frequent 1000 word bands taken from Hindmarsh (1980) and Nation (1984). The test contains a further 20 false words. Testees are required to indicate if they know each of these words. Yes responses to the false words are taken to indicate that the testee is over-estimating their knowledge and the score drawn from the Yes responses to the real words is adjusted downwards accordingly.

Procedure

The participants took the tests in class in the order: X-Lex, the generation tasks, and finally the PVLT. They were given a booklet to record all their answers. Instructions were given orally in English. There was no time limit imposed but all students completed the tasks within the 45 minutes of the class.

Analytical Procedure

The tests can be argued to have good construct validity if they can be shown to generate words across the first five 1000 words frequency bands and it is expected that frequency effects should be visible in the data produced by students. Learners should score more in the higher frequency bands than the less frequent ones. If the responses do not display this kind of frequency profile then this will undermine the potential for category generation tasks as we are using them to provide a good estimate of size.

The number of words available for selection from each of the four categories separated by the five 1000 words frequency bands (taken from the BNC/COCA lists) is shown in Table 2.

Table 2
Availability of Words in the First Five Frequency Bands Divided by Category.

	1000	2000	3000	4000	5000	Total
Animals	6	6	15	14	15	56
Clothes	16	10	4	7	14	51
Body parts	24	12	10	17	10	73
Furniture	20	39	4	10	10	83

The words produced by learners from each of these bands are compared with the number of words available for selection in each frequency band and these figures are used to generate an estimate of knowledge out of 5000. For example, if a learner were able to produce 28 of the 56 available words in the animal category then it would be assumed that this represented productive knowledge of 50% of the 5,000 most frequent words in English; a score of 2,500 words.

In testing this format's reliability the results from the four categories can be used to generate a calculation for Cronbach's Alpha. If the tests work well then the calculations generated by each test should correlate well and the Alpha score should be high.

The category generation tasks can be argued to be valid if results correlate well when compared with results from other tests of the same quality. It might be expected that they should correlate well with PVLT, which tests the same construct of productive vocabulary knowledge. They should correlate too with X-Lex, though perhaps not so well since X-Lex is, in theory, testing a slightly different construct. The tasks, if they are producing useful estimates of productive vocabulary size, should also be able to distinguish between low level learners and high level learners for example. It would be expected, too, that frequency effects should be visible. Learners should score more in the higher frequency bands than the lower frequency bands.

Finally, it might be expected that if the category generation tasks and PVLT are testing the same quality of productive vocabulary size then factor analysis and the calculation of Eigen values will confirm that a single

factor underlies the results all five tests. If receptive vocabulary knowledge is a separate and distinct construct then these calculations should show that a second factor underlies the X-Lex scores.

Findings and Discussion

Frequency Effects

Responses from the 4 generation tasks, per 1000 word frequency group and presented as an estimate of words known, are presented in Table 3 and an indication of the kind of frequency effects which emerge in the data are summarised, using figures combined from all four tasks, in Figure 1.

Table 3
Total Responses by Frequency Band

	1000	2000	3000	4000	5000
animals	316	216	168	262	229
clothing	455	292	88	127	205
furniture	501	449	210	47	40
Body parts	508	369	104	103	124

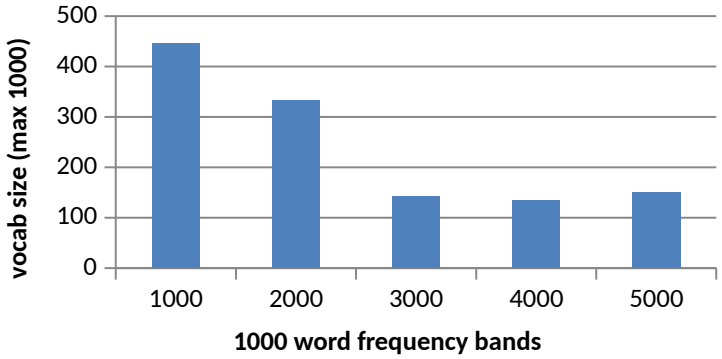

Figure 1. Frequency Effects in Productive Vocabulary

Table 3 and Figure 1 demonstrate a visible frequency effect with the bulk of learners' vocabulary knowledge lying in the most frequent 1000 and 2000 word bands. Beyond this mark and frequency effect is no longer visible. Nonetheless, productive vocabulary knowledge resembles receptive vocabulary knowledge, with the presence of the frequency profile as suggested by Ellegård (1960) and Meara (1982), and as observed in Waring (1997). The implication of this is that the category generation tasks are be capable of providing a characterisation of the scale of a learner's productive vocabulary size. Since such an estimate is similar in its calculation to a test such as X-Lex, which also draws its estimate from these frequency bands, this should allow productive and receptive vocabulary size to be meaningfully compared.

Productive Size Estimates

Productive size estimates obtained from the four category generation tasks are shown in Table 4. The four category generation tasks have demonstrated they draw words from across the first five frequency bands which means that it is possible to produce estimates of productive vocabulary size. The mean size estimates produced are in the region of about 1000 words. There is some variation here with the Furniture task producing the smallest mean estimate of 790 words, and the Clothes task the highest mean estimate of 1243 words. There are

no normalised figures for the size of productive vocabularies for learners at the levels in this study and the significance of these figures can only become apparent when compared with results from the others tests.

Table 4
Mean Word Knowledge by Category Generation Task

	Mean productive vocabulary size	SD
animals	1155.86	386.90
clothing	1243.61	380.38
furniture	790.99	243.96
body parts	1026.84	357.88

There are several reasons why the tasks used here might vary in the scores they produce. One is that the topics, taken from the literature on testing in psychology, have not been chosen with EFL testing specifically in mind. However, they are thematic areas which are typically contained in teaching texts for young and beginner learners of EFL although we have no way of knowing exactly what lexis is contained in these teaching texts nor how the treatment of this lexis may vary from one theme to another in terms of presentation and recycling. It is conceivable that these differences in measured knowledge may accurately reflect differences in the presentation of the material and this might challenge the usefulness of this approach as a quick and easily replicable method generating consistent measures for productive vocabulary size. A second is that the size of the estimate may vary according to the theme chosen for testing and not just the overall vocabulary knowledge of the learner. A third possibility is that these differences may be related to the size of the category generation task itself. Thus, the furniture category which has the largest number of words available for production produces the smallest size estimate, and clothing which has the smallest number of words available produces the largest estimate. It is also quite possible, however, that these differences are the by-product of different task forms and different administrations, where some variation in scores is inevitable even in well-constructed and regulated tests. Nation's 14,000 word multiple choice test, for example, has parallel forms which in trials, he reports (2012, p.5), produce different scores.

These differences in the means between all four category generation tasks are statistically significant, and the results of t-test and Cohen's D comparisons are given in Table 5. If parallel forms of this task consistently produce scores which are different then this challenges the validity of the testing method and the usefulness of the technique as a method for quickly and easily assessing productive vocabulary size. However, the Cohen's D calculations show that the effect size is highly variable. It is not yet clear, therefore, whether these differences *do* challenge the test's validity in this way or a simply part of the kind of variation which repeated testing produces and which Nation (2012), for example, reports in relation to receptive vocabulary size testing.

Table 5
T-test Comparisons between the 4 Category Generation Tasks

	Clothes Test		Furniture Test		Body Test	
	t-score	Cohen's D	t-score	Cohen'sD	t-score	Cohen'sD
Animal Test	2.617**	0.223	11.502**	1.13	3.511**	0.35
Clothes Test			17.836**	1.42	9.607**	0.79
Furniture Test					8.708**	0.59

Note. ** = significant at the 0.01 level

Reliability Calculations

There are moderate to good correlations between scores on the four category generation tasks. All correlations are statistically significant at the 0.01 level. The figures are shown in Table 6.

Table 6
Category Task Inter-test Correlations

	Clothes Test	Furniture Test	Body Test
Animal Test	.554**	.618**	.649**
Clothes Test		.688**	.830**
Furniture Test			.781**

Note. ** = significant at the 0.01 level

The Body parts task scores correlate particularly well with both the Furniture and the Clothes task while the Animals task scores correlate least well with the others. This observation might be connected to the number of words available for production in these tests. The Body parts and Furniture tasks have the highest number of words in the 5,000 word bands, 73 and 83 words respectively, while the Animal task has only 56 words. For comparison it might be considered that the receptive X-Lex test samples 100 words form this 5,000 word range and in the Animal and Clothing tasks there are only about half this number available for production. The reliability of the task might be influenced by the sampling rate and, as a general rule, a larger sample is likely to produce a more useful estimate. However, in this type of task a very large sample may challenge the immediate recall ability of the learner and lead to under-estimation. A thematic prompt where there are 20 words available from the 5,000 word range under examination is an achievable task but a similar task with 2,000 words is not. The impact of the potential sample size available from different themes and task is something to be investigated.

The calculation of Cronbach's Alpha using the 4 parallel forms of the productive task can be taken as an indication of the degree to which these tests measure a single construct. The Cronbach's Alpha result was .885 (N = 4). Notwithstanding potential difficulties with individual category tasks and their sampling rate, the score of .885 is good and can be taken as confirmation that these tasks can produce results which are both reliable and consistent.

Productive Scores by Level

Mean productive vocabulary size scores generated by each of the four category generation tasks, divided by the level of the students, are shown in Table 7.

Table 7
Mean Productive Vocabulary Size Scores by Level

Level	animals		clothes		furniture		body parts	
	mean	sd	mean	sd	Mean	sd	mean	Sd
Beginner	910	343	745	200	606	176	940	209
Intermediate	1102	298	995	223	754	114	1185	195
Advanced	1461	265	1357	289	1019	182	1616	296

The productive size scores generated by all four tasks increase with the level of the students as is expected. The advanced group of learners produce in each task, on average, more words from the 5,000 word frequency ranges, than the intermediate level students who, in turn, can produce more words on average than the students at the beginner level. An ANOVA confirms that this relationship is statistically significant and the results are shown in Table 8. Tukey tests confirm that there are statistically significant differences between the means at all levels in all tests. The ability of these tasks to discriminate meaningfully between learners at different levels of knowledge and performance supports the construct behind the test and suggests this technique is valid.

Table 8
ANOVA Scores from the Category Generation Tasks

test	degrees of freedom	F	Sig
animals	between groups 2	28.439	< .001
	within groups 89		
clothes	between groups 2	55.648	< .001
	within groups 89		
furniture	between groups 2	54.277	< .001
	within groups 89		
body parts	between groups 2	68.634	< .001
	within groups 89		

PVLT And X-Lex Scores And Inter-test Correlations

If the new test form is to demonstrate concurrent validity then test scores should correlate with scores from others tests of the same or related constructs. The new tests should correlate acceptably with PVLT, which is a test ostensibly of exactly the same construct, and should correlate too with X-Lex scores, which tests a closely related construct. PVLT mean scores per level and the overall means are shown in Table 9 and X-Lex mean scores per level and the overall means are shown in Table 10.

Table 9
PVLT Scores Divided by Level

	n	mean	Sd
Beginner	36	982	745
Intermediate	23	910	717
Advanced	33	2124	1348
Total	92	1338	1138

Table 10
X-Lex Scores Divided by Level

	n	mean	Sd
Beginner	36	3084	845
Intermediate	23	2737	567
Advanced	33	3685	790
Total	92	3213	847

Correlations between PVLT and X-Lex scores, and the scores on the four category generation tasks are shown in Table 11.

Table 11
Correlations between Category Generation Task Scores and PVLT and X-Lex Scores

	PVLT	X-Lex
Animals test	0.494**	0.362**
Body parts test	0.408**	0.424**
Furniture test	0.344**	0.324**
Clothes test	0.481**	0.353**

Both PVLT scores and X-Lex scores indicate, broadly, that the vocabulary size of the learners increases with level as might be expected and this is confirmed by ANOVAs (PVLT F (2,89) = 14. 539, sig<.001, X-Lex F

= (2,89) = 11. 237, sig<.001). Tukey tests, however, indicate that neither test is able to produce a statistically significant difference in the means between the Beginner and Intermediate students. The category generation tasks were capable of doing this and one interpretation of this is that the category generation tasks are better able to distinguish levels of knowledge among lower level learners than the other tests. PVLT produces an estimate of size which is slightly larger than the estimates produced by the category generation tasks. An analysis of variance used to calculate effect size suggests a moderately large effect size but this result is not statistically significant (F(89,2)=5.312, sig=.171). This may be a product of the different methodologies and knowledge being accessed. PVLT provides quite extensive context and a letter cues for each test word where the category generation tasks so not. Or it may be an outcome of the formula for turning PVLT scores into a size estimate where not all frequency bands are tested and knowledge in these missing bands has to be inferred from knowledge elsewhere. The difference between the means for the PVLT and the largest scoring category task, Clothes, is not statistically significant. The difference between the means for PVLT and Animals is significant only at the .05 level (t = 2.077, sig = .041). There are significant differences between PVLT and the means for the other two tests (Furniture t = 3.138, sig = .002, Body parts t = 5.177, sig < .001).

X-Lex produces a larger estimate of vocabulary size than either the category generation tasks or PVLT. X-Lex is a receptive vocabulary size test and it is expected that receptive size estimates will be larger than productive size estimates. An analysis of variance used to calculate effect size produces a result that is not statistically significant (F(89,2)=1.016, sig=.622). In a review of the literature in this area Milton (2009), Nation (1990) and Schmitt (2000) report that the difference between these scores varies but that, typically, receptive sizes are about double that of productive sizes. In this study the scores suggest that the productive size estimates are between one third and a half of the size of the receptive estimates and the relationship is summarised in Figure 2. This figure suggests that while the five productive sizes mean scores can be distinguished statistically, they are of similar scale and in the right kind of proportion in relation to receptive vocabulary size. It may be that refining the category generation tasks can make them perform more consistently in producing more similar size estimates.

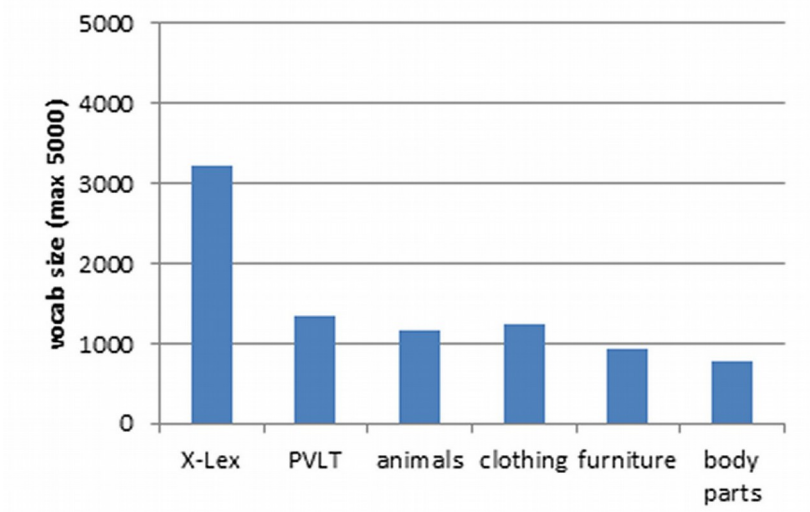

Figure 2. Comparison of Receptive and Productive Vocabulary Size Estimates

Factor Analysis

Since PVLT and the four category generation tasks are all designed to access productive vocabulary knowledge and produce estimates of productive size, it is expected that factor analysis should reveal a single factor underlying the scores. Factor analysis and the calculation of Eigen values allows this to be investigated. The scree plot (Figure 3) and component matrix (Table 12) suggest that this is the case. The scree plot identifies only one

component with a score above 1. The component matrix indicates that the four category generation tasks all correlate well with this factor while the correlation produced with PVLT is smaller but still satisfactory.

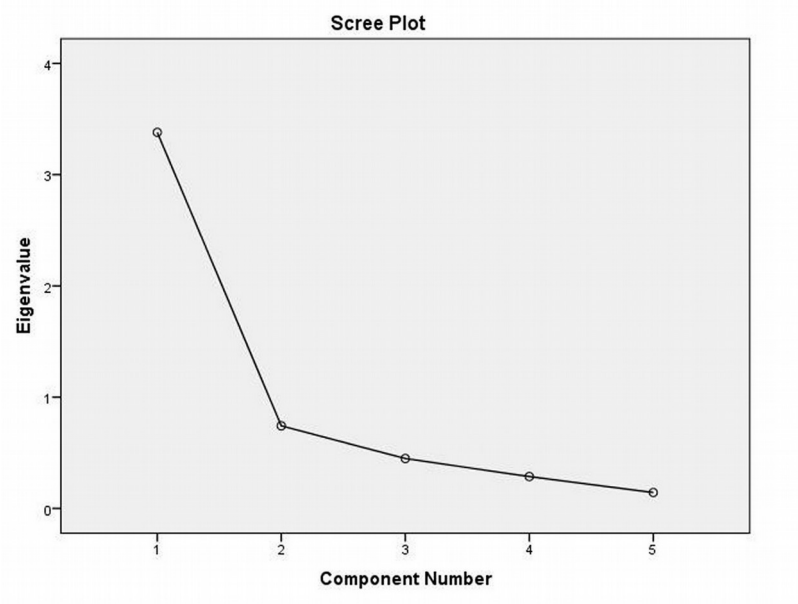

Figure 3. Scree Plot for Productive Vocabulary Size Tests

Table 12
Component Matrix for Productive Vocabulary Size Tests

	Component 1
Animals	.806
Clothing	.865
Furniture	.854
Body parts	.928
PVLT	.626

It is expected too, that when the five productive tests and X-Lex are compared that more than one factor should be visible since X-Lex is designed to access a different construct from the others and that receptive knowledge is considered to be qualitatively and quantitatively different from productive knowledge. It is not clear from the factor analysis that this is visible. The scree plot (Figure 4) and component matrix (Table 13) suggest that a single factor underlies the scores in all six tests even if X-Lex, like PVLT, correlates less well with this single factor than the category tasks. The implication of this is that receptive and productive knowledge scores are all, largely, explained by just one factor. We presume this is vocabulary size but it could be other things. It could be a general vocabulary knowledge factor or it could be a something non-linguistic like intelligence.

It is fashionable to think of vocabulary as multidimensional but these results suggest that one of the oldest divisions of vocabulary knowledge, receptive and productive knowledge, may not be quite the division that is thought. Of course, receptive and productive knowledge cannot be completely unrelated. A condition of having a large productive vocabulary knowledge is having a large receptive vocabulary knowledge; it is presumably impossible to produce meaningfully words in a foreign language that are not even recognised as words. In principle, it should be possible for the reverse to be true and for a large number of words to be recognised even if knowledge is so limited that they cannot be activated and used. However, our interpretation of the factor

analysis, and correlations between the productive and respective tests, is that in practice productive knowledge tends to grow with receptive knowledge. Co-linearity is a feature of the studies which compare vocabulary size with automaticity in production (Schoonen, 2010).

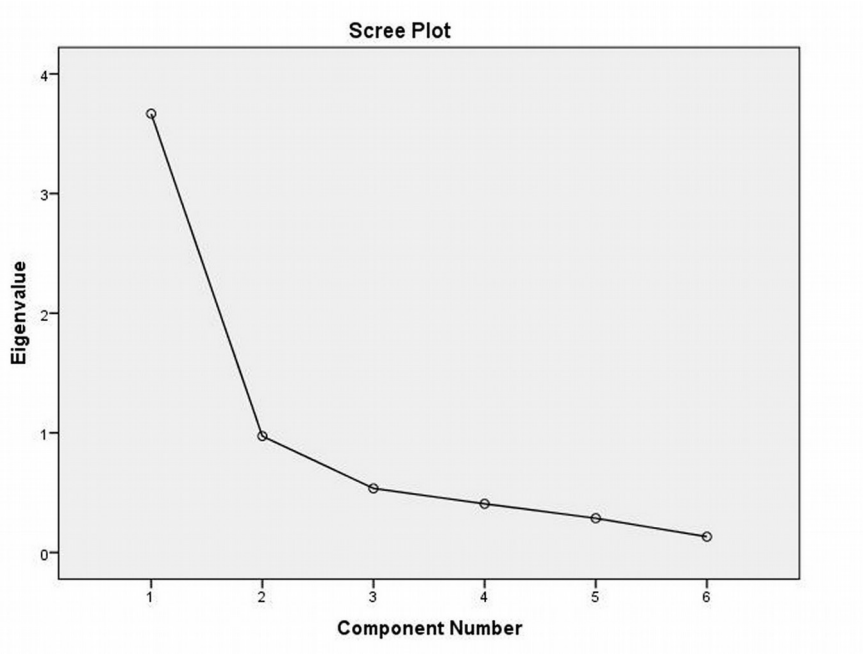

Figure 4. Scree Plot for All Vocabulary Size Tests

Table 13
Component Matrix All Vocabulary Size Tests

	Component 1
Animals	.794
Clothing	.856
Furniture	.828
Body parts	.902
PVLT	.669
X-Lex	.599

Conclusions

What can we conclude from this? It is possible to make a case that the category generation task is, potentially, a useful test format which can measure, and put a size on, productive vocabulary knowledge. The tests have proven reliable and, in certain ways, valid. The category generation task triggers learners at all levels to produce a large number of words with minimum direction or interference from the teacher or a text. It is able to target a predictable range of words in the frequent vocabulary bands so that a workable estimate of productive knowledge can be formed, and these estimates correlate reasonably well with each other so the Alpha score is high. It distinguishes between low, intermediate and high level learners well, arguably rather better than PVLT or X-Lex. It correlates, although modestly, with other tests of productive and receptive vocabulary knowledge, and this suggests that teaching effects may not be significantly affecting the ability of the technique to make a

good estimate of size. It also produces scores, consistently, which are smaller than receptive vocabulary size which makes sense. It is a very easy format that requires very little adaptation to work across learners from different language backgrounds, and it may be particularly useful in assessing knowledge among very low level learners. This type of test for productive vocabulary size seems to have potential, therefore, but this study has raised questions about the use of the technique and the estimates it creates which need to be investigated more thoroughly.

One is that the separate scores from the different category generation tests and the PVLT all produce different mean scores and, with one exception, the differences are sufficiently great to be statistically significant. Parallel forms which give a stable size estimate are necessary if the test is to perform like the receptive tests of vocabulary size and be capable of being used as a standard test in this area. Nonetheless, the scores that are produced are all about one third the estimate of receptive vocabulary knowledge and that ties in with other studies in the literature which compare receptive and productive vocabulary knowledge. It has already been noted that parallel test forms rarely produce identical scores, but what should be made of the scale of variation seen here is, as yet, unclear. As Meara (2009) points out, the words produced for assessments in productive tasks are dependent on the task, the genre and the prompt itself so, perhaps, a range of scores is what we should be seeing if students are responding to a range of tasks even if their vocabulary remains unchanged. The construct of productive vocabulary could usefully benefit from a more precise specification to help us work through these difficulties.

It has to be noted too that this is just one study based on learners with one language background and in one country. It would make sense to repeat this form of testing on other learners with different learning and language backgrounds as a check to see that the technique is applicable beyond learners in Iran.

There are issues too with the prompts used in this study which are a small group of prompts drawn from the psychology literature. These prompts were chosen not least because they are also areas typically covered in young learner syllabuses. But this may make the scores they produce potentially misleading since words drawn this way may also challenge the underlying idea that a good estimate of size is made by using a random sample of words across the frequency bands. A sample that draws on the subject areas that we know that learners have covered is not a random sample. The effect of such a choice of prompt also needs to be clarified although it is not clear from this study that any effect that does exist is very great.

The sampling across the frequency bands, produced by these prompts, produces a workable selection, from which an estimate can be made. But it is notable that the selections this produced are of different sizes and not evenly spread across the frequency bands. The effect on the estimate this produces will need to be measured and appraised. Given the issues which may surround the size of the potential sample a thematic prompt can produce, it would also make sense to repeat this work with other prompts. It would make sense to investigate prompts capable of producing larger samples in order to test the effect of this on the size of the estimate. Large prompts seem likely to produce smaller estimates. It would be useful to know at what levels the estimates appear less than useful. It would make sense, too, to investigate prompts capable of producing better and more equally sized samples. This would seem likely to help control for the variation in scores produced by the four tests used in this study. This would require the use of themes other than the four used in this study which were, in any case taken from psychology. If the methodology is to prove useful in EFL then a wider variety of themes, perhaps more directly applicable to EFL testing, might be appropriate. It might even be useful to test the use of other prompts such as letters of the alphabet rather than thematic cues although in the psychology literature, these appear to work rather differently.

Finally the factor analysis is raising an unexpected question since it appears that productive and receptive vocabulary knowledge used here are not the separate constructs as they are generally portrayed but are all tapping into a single factor which may be some general vocabulary knowledge or size. Maybe that should not be surprising since the various dimensions of vocabulary knowledge ought to be connected. The ability to produce a word has as a precondition that the word is known receptively, so it follows that a large productive vocabulary knowledge must be associated with a large receptive score. High productive and low receptive scores ought to be

impossible if the construct of the lexicon is as we understand it, and the tests we use to access knowledge are working tolerably well. The opposite may be potentially true, where a high receptive knowledge might be associated with a small productive knowledge, but it is hard to imagine the circumstances of teaching and learning that might produce a very highly disparate set of scores. The common acceptance of the idea of multi-dimensionality in vocabulary knowledge and the need for multiple testing, therefore, should not blind us the way these dimensions necessarily interconnect. Our interpretation of the factor analysis in this study is that for most practical purposes, the need for multiple testing in vocabulary is probably not as important as is thought. Multiple testing may be useful in the research community but it seems as though for most practical purposes a single well-constructed test is likely to give a good impression of all aspects of vocabulary knowledge.

This study suggests that in its present form the test would be useful in schools in order to generate an estimate of size so learners can be ranked or compared on their productive knowledge. Where a productive test in particular is wanted, this will likely work well. However, it is not yet in a state where parallel forms can be generated and a stable estimate of size produced and used, as in receptive vocabulary size tests, for use in research or to link with other factors of language performance like exam performance.

References

Cobb, T. (2014). http://www.lextutor.ca/. (accessed 31st August 2014).

David, A. (2008). Vocabulary breadth in French L2 learners. *Language Learning Journal*, 36(2), 167-180.

Ellegård, A. (1960). Estimating vocabulary size. In *Word*, 16, 1960, 219-244.

Ellis, N. C. (2002a). 'Frequency effects in language processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in second language acquisition*, *24*(02), 143-188.

Ellis, N. C., (2002b). Reflections on frequency effects in language processing. *Studies in second language acquisition*, *24*(02), 297-339.

Fitzpatrick T. and Milton J. (2014). Reconstructing vocabulary knowledge. In Milton, J. and Fitzpatrick, T. (eds.) *Dimensions of Vocabulary Knowledge*(pp. 173-177). Basingstoke: Palgrave.

Hindmarsh, R. (1980). *Cambridge English Lexicon.* Cambridge: Cambridge University Press.

Izura, C., Hernández-Muňos, N. and Ellis, A. (2005) Cognitive norms for 500 Spanish words in five semantic categories. *Behavior Research Methods,* 37(3), 385-397.

Laufer, B. (1989). What percentage of text is essential for comprehension? In Lauren, C. and Nordman, M. (eds.) *Special Language; from Humans Thinking to Thinking Machines* (pp. 316-323). Cleveden: Multilingual Matters.

Laufer, B. & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing,* 16(1), 33-51.

Long, M. and Richards, J (2007). Series Editors' Preface. In Daller, H., Milton, J. and Treffers-Daller, J. *Modelling and Assessing Vocabulary Knowledge* (pp. xii-xiii). Cambridge: Cambridge University Press.

Meara, P. (1982). Word association in a foreign language: a report on the Birkbeck vocabulary project. *Nottingham Linguistic Circular,* 11, 29-37.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt, and M. McCarthy, (Eds.) *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge; Cambridge University Press, 109-121.

Meara, P. (2009). *Connected Words: Word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins.

Meara, P. and Bell, H. (2001). P-Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect* 16(3), 323-37.

Meara, P. and Milton, J. (2003). *The Swansea Levels Test.* Newbury: Express.

Meara, P. M. and Miralpeix, I. (2008). Vocabulary Size Estimations: V_Size 41st Annual Meeting of the British Association for Applied Linguistics (BAAL). Swansea, UK.

McKinney, K L (2009). *Lexical Errors Produced During Category Generation Tasks by Bilingual Adults and Bilingual Typically Developing and Language-Impaired Seven to Nine-Year-Old Children.* Unpublished MA thesis The University of Texas at Austin.

Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In *Daller, H., Milton, J. and Treffers-Daller, J. (eds.) Modelling and Assessing Vocabulary Knowledge* (pp. 47-58). Cambridge: Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In Vedder, I. Bartning, I. & Martin, M. (eds.) *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (pp. 211-232). Second Language Acquisition and Testing in Europe Monograph Series 1.

Milton, J. & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In Richards, B., Daller, M., Malvern, D., Meara, P., Milton, J. & Treffers-Daller, J. (eds.) *Vocabulary Studies in First and Second Language Acquisition* (pp. 194-21). Basingstoke: Palgrave.

Milton, J. & Riordan, O. (2006). Level and script effects in the phonological and orthographic vocabulary size of Arabic and Farsi speakers. In Davidson, P., Coombe, C., Lloyd, D. and Palfreyman, D. (eds) *Teaching and Learning Vocabulary in Another Language* (pp. 122-133). UAE: TESOL Arabia.

Milton J., Wade, J. & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In Chacón-Beltrán, R., Abello-Contesse, C. & Torreblanca-López, M. (eds.) *Further insights into non-native vocabulary teaching and learning* (pp. 83-98). Bristol: Multilingual Matters.

Nation, I.S.P. (ed) (1984). *Vocabulary Lists: words, affixes and stems.* English University of Wellington, New Zealand: English Language Institute.

Nation, I S P (1990). *Teaching and Learning Vocabulary*. Boston: Heinle and Heinle.

Nation, I.S.P. (2001). Vocabulary Levels Test. In Nation, I.S.P. (2001) *Learning Vocabulary in Another Language* (pp. 416-424). Cambridge: Cambridge University Press.

Nation, I.S.P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In Daller, H., Milton, J. & Treffers-Daller, J. (Eds.) *Modelling and Assessing Vocabulary Knowledge* (pp. 33-43). Cambridge: Cambridge University Press.

Nation, I.S.P. (2012). Vocabulary Size Test. instructions a t http://www.victoria.ac.nz/lals/about/staff/paul-nation (accessed 31st August 2015).

Palmer, H.E. (1921.) *The Principles of Language Study.* London: Harrap.

Richards, B.J., & Malvern, D.D. (2007) Validity and threats to the validity of vocabulary measurement. In Daller, H., Milton, J. & Treffers-Daller, J. (Eds.) *Modelling and Assessing Vocabulary Knowledge* (pp. 79-92). Cambridge: Cambridge University Press.

Schmitt, N (2000). Vocabulary in Language Teaching. Cambridge: Cambridge University Press.

Schoonen, R. (2010). The development of lexical proficiency knowledge and skill. Paper presented at the Copenhagen Symposium on *Approaches to the Lexicon*, Copenhagen Business School on 8-10 December 2010. Accessed at https://conference.cbs.dk/index.php/lexicon/lexicon/schedConf/presentations on 03.03.2011.

Stæhr, L.S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, 13(2), 138-163.

van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In Daller, H., Milton, J. & Treffers-Daller, J. (Eds.) *Modelling and Assessing Vocabulary Knowledge* (pp. 95-115). Cambridge: Cambridge University Press.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics* (2001) 22, 217-234.

Waring, R (1997). Comparison of the receptive and productive vocabulary knowledge of some second language learners. *Immaculata; The Occasional Papers of Notre Dame Seishin University*. 1997, 94-114.

Webb, S. (2005). The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, 27, 33-52.

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, 20(2), 232-245.

About the Authors

Shadan Roghani is a qualified teacher who has worked for 14 years in secondary schools. She obtained her Master's degree in TEFL from Swansea University in 2013. She is currently a PhD student in Swansea investigating the topic 'Using Lexical Generation Tasks in Testing Productive Vocabulary Size.'

James Milton is Professor of Applied Linguistics at Swansea University, UK. He worked in Nigeria and in Libya before coming to Swansea in 1985. A long-term interest in measuring lexical breadth and establishing normative data for learning has produced extensive publications including *Measuring Second Language Vocabulary Acquisition* (Multilingual Matters, 2009).