Words as species: an alternative approach to estimating productive vocabulary size.

PM Meara *Swansea University*
and JC Olmos Alcoy *University of Dundee.*

1: Introduction

This paper is the fourth in a series of studies in which we have tried to address the question of estimating productive vocabulary size in L2 speakers. The basic distinction between active and passive vocabulary is a staple idea that is widely taken for granted in introductory books on vocabulary acquisition, and in instructional texts designed to teach vocabularies. Some writers, for example, go so far as to list vocabulary items which need to be acquired productively, and others where it is sufficient for learners to be able to recognise them passively. Despite the fact that very many researchers have written about this topic, (Melka 1997, Melka Teichroew 1982, 1989, Laufer 1998), the idea of productive vocabulary remains a fundamentally elusive one, and it has proved surprisingly difficult to develop tests of productive vocabulary size with any degree of face validity. The test most widely used in the research literature is probably Laufer and Nation's Productive Levels Test (Laufer and Nation 1999), a simple adaptation of Nation's very successful Levels Test, which is widely used to estimate receptive vocabulary size (Nation 1990). Laufer has used these two tests to make some very interesting claims about the relationship receptive and productive vocabulary, and how these two facets of vocabulary knowledge develop at different rates (Laufer 1998). However, the data provided by the Productive Levels Test is far from staightforward, and in our view it is worth while looking at alternative approaches to estimating productive vocabulary size.

In our previous research, we have developed three main ideas, which we think will allow us to “triangulate” the idea of productive vocabulary size. The first of these ideas involved moving away from using written texts as the raw data for research on productive vocabulary size. Meara and Fitzpatrick (2000) argued that ordinary texts generated by learners tended to contain very large numbers of highly frequent words, and very few infrequent words which were true indicators of a large productive vocabulary. They tried to get round this problem by getting learners to generate “texts” derived from a set of word association tests. These data typically consisted of infrequent L2 words, and Meara and Fitzpatrick argued that they provided a better picture of the scope of the testee's productive vocabulary than other, more traditional test types did. Unfortunately, it was not obvious how the scores provided by their Lex30 test could be converted into proper estimates of vocabulary size, which could be used to address the questions raised by Laufer.

In our second approach to estimating productive vocabulary, Meara and Bell (2001) returned to using texts generated by L2 writers, and attempted develop what they called an “extrinsic measure of vocabulary richness”. They analysed sets of short texts produced by L2 learners, and for each text generated a curve which described the incidence of “unusual” words in short segments of text. They then showed that these curves could be summarised in terms of a single parameter, λ, and argued that this parameter might be related to overall productive vocabulary size. This approach successfully distinguished between learners of English at different proficiency levels, but as with the Lex30 test, Meara and Bell were not able to establish a direct, quantifiable relationship between λ and overall productive vocabulary size.

In our third approach, Meara and Miralpeix (2007) attempted to estimate productive vocabulary directly by looking at the frequency distribution of words used by L2 writers, and comparing these profiles to a set of theoretical profiles derived from Zipf's law (Zipf 1935). Meara and Miralpeix argued that it might be possible to estimate a learner's productive vocabulary size by identifying a theoretical vocabulary profile which closely matched the actual data produced by the learner. This general approach is solid enough to distinguish between advanced and less advanced learners. More importantly, this approach actually allows us to quantify the productive vocabulary that seems to be behind a particular text. For example, it allows us to make statements like “the text in Table 1 implies a productive vocabulary of around 6400 words.” This is a significant advance, which opens up a number of promising avenues of research, but it rests on a number of assumptions about the way L2 learners acquire words, which may not be fully justified.

Table 1: V-Size estimates that this text was generated by a speaker with a productive vocabulary of at least 6400 words.

Once upon a time there was a dark and lonely wood, where three bears lived. The bears lived in a small cottage at the end of a dark and lonely road, where few people ever strayed. The bears liked it a lot. They did not get many visitors, but that was fine. The rest of the time they kept to themselves, and went about their business in a calm and peaceful way.

Father Bear was the one who liked the dark and lonely bit best. He was a philosopher by nature, who loved to read dark and lonely poetry written in the dead of Winter by Scandinavian poets who also lived in dark and lonely woods, and generally suffered from Angst. Mother Bear didn't have much time for Angst. She was practical and organised, and liked the dark and lonely wood because nothing ever happened there to disturb her domestic routine. Yes, it would have been nice if Father Bear did a bit more of the cooking and cleaning, and yes, it would have been nice if Tesco had a branch at the edge of the wood, but it was better than having noisy neighbours who bothered you all the time. Baby Bear still hadn't decided if he liked the dark and lonely wood or not. It was scary at night, and it was easy to get lost in the wood if you forgot to leave your marks on the trees where the paths split. But Baby Bear had been to the town once too, and he definitely did not like it. Not one bit.

For these reasons, we have also been pursuing other approaches to estimating vocabulary size. Our hope is that these different approaches will all turn out to provide answers which are broadly similar, and if we could achieve this, then it might be possible to develop a reliable, practical test of productive vocabulary size, which would allow us to address Laufer's questions in a principled kind of way.

This paper comes at this issue in a way which is rather different from the approaches we have developed in our previous work.

2: Estimating population sizes in the field.

The main problem with estimating productive vocabulary size is that it is extremely difficult to get all the data that we need from our experimental subjects. If we were dealing with learners with very small vocabularies, then it might be possible to devise a set of tests which assessed whether our learners could produce each of the words in a short list of words that we are interested in. In practice, however, this only works where we are dealing with very small numbers of words. In real testing situations, it would be logistically impossible to test the entire vocabulary of a learner who has more than a very elementary vocabulary. Threshold Level Spanish (Slagter 1979), for example, comprises a lexicon of around 1500 words, but speakers at this level have only a very limited level of competence in Spanish. Testing vocabulary exhaustively at this level is difficult, and just about feasible with very co-operative subjects. Testing the vocabulary of more advanced Subjects becomes increasingly difficult as their vocabulary grows. Consequently, if we want to test the vocabularies of even moderatly advanced students, we have no option but to resort to sampling methods, and to extrapolate from the results we get when we test a small number of words. Obviously, the trick here lies in devising a sampling method which is appropriate and transparent. We may not be able to get L2 learners to produce for us **all** the words that they know, but we might be able to develop a testing methodology which allows us to extrapolate meaningfully from the words that we can elicit.

This problem is not unique to linguistics. Analogous problems also occur in other areas of study, and are particularly important in ecology, where we want to count the number of animals in a given habitat area. A typical problem of this sort is when we want to estimate the number of deer inhabiting a forest, the number of elephants occupying a national park, or the number of cockroaches infesting a hotel. Simply counting the animals is not staightforward: the animals are not co-operative, and do not line up in a way which allows us to count them reliably. This makes it notoriously difficult to make good estimates of animal populations - a problem which can have serious consequences if we are trying to manage the population, and control the number of animals which a particular environment can provide for, or as in the case of the cockroaches, to eliminate them altogether.

Ecologists have developed a number of methods which allow them to get round this problem. All of these methods rely on capturing a small number of animals, and then extrapolating this basic count to an estimate of the actual number of animals that could have been caught. The basic approach is known as the capture-recapture methodology, first developed by Petersen (1896), and further developed by Lincoln (1930). In this approach, we first develop a way of “capturing” the animals we are interested in, and standardise it. Suppose, for example, that we want to count the number of caterpillars in a cabbage patch. We could identify a set of 10 cabbages, randomly distributed across the patch, and count the number of caterpillars on each of these plants. Now let us mark these caterpillars in some way, perhaps by putting a dot of paint on them. Next day, we carry out the same counting exercise, enumerating the caterpillars that we find on the same set of 10 cabbages. This gives us three numbers: we have N, the number of caterpillars captured on Day 1; M, the number of caterpillars captured on Day 2; and X, the number of caterpillars which were captured on both occasions. Petersen argued that it was possible to extrapolate from these figures to the total number of caterpillars in the patch. Petersen's estimate is calculated as follows:

$$E = (N * M) / X \qquad \text{eq. 1}$$

i.e. Petersen's estimate of the size of the caterpillar population of the patch is the product of the two separate counts divided by the number of caterpillars counted on both occasions. A simple example will make this idea more concrete. Suppose that on day 1 we count 100 caterpillars on our ten cabbages, and we mark them all. On day 2, we find 60 caterpillars, twenty of which were also noted on day 1. Petersen's estimate of the number of caterpillars in the patch would be:

$$E = (100 * 60) / 20 = 6000/20 = 300.$$

There are a number of points to make about this estimate. Firstly, the estimate is quite a lot larger than the totals counted on either of the two data collection times. Secondly, it assumes that the way we counted the caterpillars was a reasonable one, one which gave us a good chance of capturing the caterpillars we want to count, and that the 10 cabbages we have selected “represent” in some way the entire cabbage patch. Thirdly, the mathematics only works in a straightforward way if we assume that the two collection times are equivalent, and if each animal has an equal chance of being counted on both collection times. The population of caterpillars needs to be constant from Day1 to Day2 - if half our caterpillars were eaten by hedgehogs, or turned into butterflies overnight, then Petersen's model would simply not apply. Finally, we are assuming that the data collection on Day2 is “equivalent” to the data collection on Day1, and so on. If these assumptions do not hold, then the model will not work, but if the assumptions are broadly correct, then these two capture events allow us to make a rough estimate of the number of caterpillars in the patch, even though we are not able to count every single one of them.

Petersen's method has been widely used in ecological studies, where researchers have been interested in estimating the size of elusive animal populations, and it turns out to be surprisingly accurate and reliable. Seber (1982) and (1986) provide a number of examples of how the method has been used in practice.

The question we ask in this paper is whether it might be possible to adapt this approach to making estimates about productive vocabulary size? At first, it seems unlikely that this ecological approach would provide a good analogy for what happens with words. Words are not animals, and their characteristics are very unlike those of caterpillars or elephants. Indeed, you could argue that words are not entities at all - rather they are processes or events, which need to be counted in ways which are different from the ways we use to count objects. Nevertheless, there seems to be a case for exploring this idea a little further, before we reject it out of hand.

One immediate objection is that the method as we have described it so far seems to work well for counting individual animals, but when we count words we are not really interested in how many exemplars of a single word we find. More usually we are interested in how many different **word types** we can identify in a text. This is more like counting the number of different caterpillar species we find in our cabbage patch, rather than the number of actual caterpillars. For example, suppose that our first data collection event delivers 10 types of caterpillar, and we make a record of these 10 types. If our second data collection delivers 12 types of caterpillar, of which 8 were previously recorded, then Petersen's estimate of the number of caterpillar types in our cabbage patch is:

$$E= (10*12)/8 = 120/8 = 15.$$

This approach to measuring the number of different species in a site uses essentially the same mathematics as the earlier example, but counts the number of different caterpillar **types**, rather than the number of different caterpillar **tokens**. This shift in focus seems to us to be an interesting one, which readily leads into better analogies with words. The main difficulty is that while it is relatively easy to devise traps or hides which allow us to observe animals and count species, it is much less obvious how one goes about building equivalent traps for words. However, as a first stab, in this paper we are going to assume that a good way of trapping words is to get speakers to write short essays. Some of the problems with this assumption will be given further consideration in Section 5.

3: METHODOLOGY

subjects

24 subjects took part in this study. All of them were learning Spanish at the University of Dundee. 11 of the subjects were low intermediate level, while the remaining 13 subjects were considered by their teacher to be “advanced”. These Ss were all L1_English speakers.

data collection

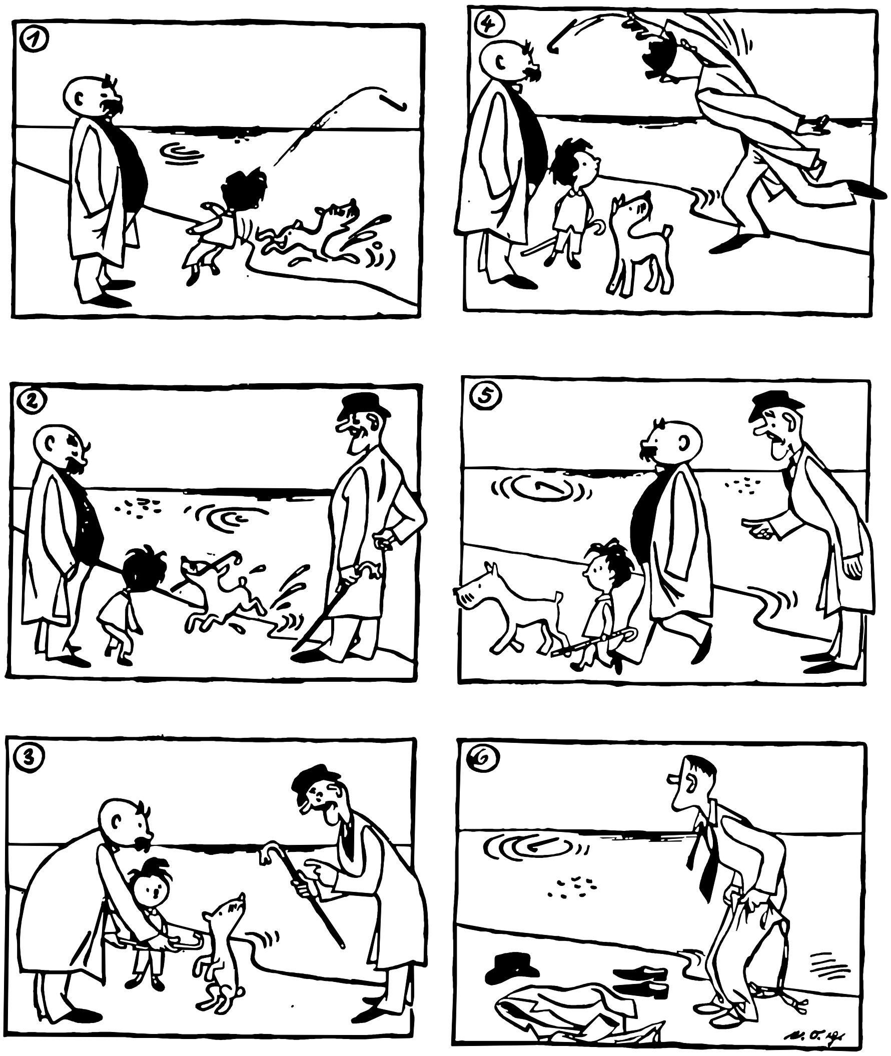
The 24 Ss were asked to write a description of a cartoon story. The story (reproduced in Figure 1) consisted of six pictures. Ss were given 30 minutes to write their accounts, and during this time they were not allowed to use dictionaries, or to confer with their colleagues. This same procedure was repeated a week later, when Ss were asked to write a second description of the same cartoon story. In both data collection events, Ss write their stories by hand. The hand-written stories were then collected and transcribed into machine readable format for further analysis. Table 2 illustrates the kind of material that was generated by this task.

Table 2: a sample text elicited by Figure 1.

> Hay un hombre y un niño cerca de un rio y el hombre esta mirando el niño, el niño esta jugando con el perro y se tira un ayuda de andar de madera en el rio.
>
> El perro llega del agua con el ayuda de andar de madera y aparece un hombre, alto y delgado, con un ayuda de andar de madera, tiene la ropa muy formal y un sombrero. Este hombre nuevo esta mirando el niño y el pero con un sonrisa.
>
> El hombre original y el niño toman el madera del perro y el hombre formal empieza a enseñar a el perro su ayuda de andar de madera. El perro, el hombre original y el niño estan

> mirando a el hombre formal.
>
> El hombre formal empieza a tirar su ayuda de andar de madera en el rio, con gran fuerza, se usa todo su cuerpo para tirar y el madera va muy, muy lejos en el rio. El hombre original, el niño y el perro estan mirando, sin movimiento, a el hombre formal.
>
> Ahora el ayuda de andar de madera esta en el rio, muy lejos y el hombre original, el niño y el perro estan andando fuera, ya tienen todos sus posesiones y estan contentas. El hombre formal esta muy discontenta, su madera esta lejos y en el rio. El hombre formal pregunta a el perro, el hombre y el niño para que queden y el perro trae el madera del rio.
>
> Ahora el hombre formal esta solo y esta mirando el ayuda de andar de madera pero al mismo tiempo esta sacando todo su ropa para que nade a su madera. Su sombrero, zapatos, chaqueta y camiseta estan en el suelo y ahora mismo el hombre formal esta sacando sus pantalones.

Figure 1: The cartoon story used to elicit the L2 texts



Because the students are fairly low level, some leniency was used in the transcriptions. Orthographic errors were corrected, and grammatical errors were ignored. The transcriptions were submitted to a computer program which reported the number of word tokens and the number of word types for each text. In calculating these figures, a number of ad hoc decisions had to be made about how to handle different word forms in Spanish. Noun and Adjective forms which varied in number or gender were considered as exemplars of a single word type. So, **guapa, guapas** and **guapos** were considered to be variants of a single type **guapo.** For verbs, the same principle applied, except that verbs in the same tense were considered to be examples of a single type, while irregular forms and different tenses were counted as separate types. Thus, **soy, eres, es** would count as three tokens of the word type **ser**, while **fuiste** and **seremos** would count as additional word types. In fixed expressions such as **por una parte, desde luego,** or **por otro lado** each word was counted separately. English words were not included in the transcripts, and words that were so badly spelled that they were unrecognisable were also deleted from the transcripts.

4: RESULTS

Table 3 shows the mean number of word tokens that the two groups generated for each of the two collection times. The table suggests that the texts of the advanced group tend to be longer than those of the less advanced group, but there is a striking difference between the text lengths of the intermediate Group at T1 and T2. An analysis of variance in which the main effects were Group and Test Time confirmed that there was a significant Group effect [F(1,22=24.19, p<.001]. Paired t-tests confirmed that the number of tokens generated by the Intermediate group was significantly greater for the second narrative than for the first, (t=3.37, p<.01 with 10 df) though the Group by Test Time interaction is not significant.

This data is fairly straightforward to interpret. The difference between the groups is what we would have expected, since text length is generally a good indicator of L2 proficiency. The significant test effect for the Intermediate Group is more difficult to interpret, and will be discussed further in section 5.

Table 3: Mean number of word tokens in two narrative description tasks.

		Narrative 1	*Narrative 2*	*Combined*
Gp Advanced	**Mn**	190.23	199.15	389.38
	sd	*48.72*	*63.65*	*59.81*
Gp Intermediate	**Mn**	99.18	133.63	232.81
	sd	*27.16*	*40.28*	*89.94*

Table 4 shows a more complex data set which records for each Subject the number of different **word types** they produced in each of the data collections, along with the number of word types which occurred in both narratives. The data suggest that the advanced group produces more word types than the intermediate group. It also suggests that for the advanced group the two tasks broadly elicited the same number of types, while for the intermediate level group, the number of types elicited in the second data collection was significantly greater than the number of types elicited in the first data collection. A t-test confirmed that this difference was significant for the intermediate group (t=2.83, p=0.017). An Analysis of Variance in which the main effects were Group and Test confirmed that there was a significant overall difference between the advanced group and the intermediate group, but failed to show a signficant test effect, or any significant interaction between Group and Test.

Table 4: Mean number of word types in two narrative description tasks.

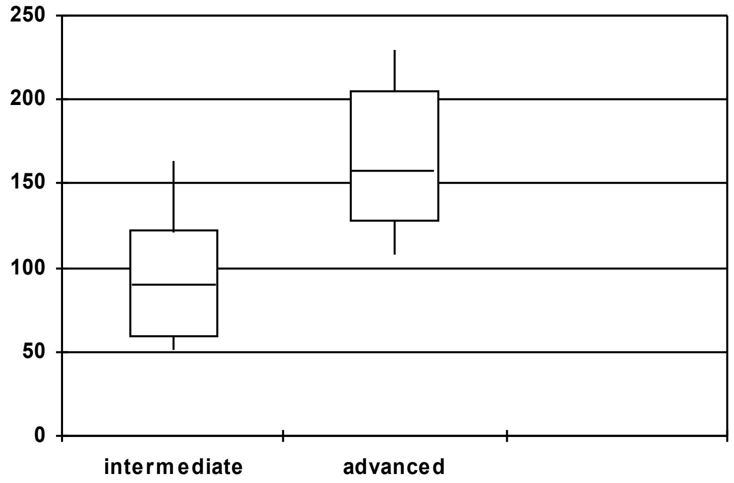
		Narrative 1	*Narrative 2*	*Common types*
Gp Advanced	**Mn**	72.91	73.73	33.55
	sd	*17.00*	*19.09*	*9.11*
Gp Intermediate	**Mn**	43.36	52.36	25.82
	sd	8.89	*15.09*	*6.91*

For each Subject, the raw numbers of types figures were plugged into the Petersen estimate formula, and the estimates generated in this way are reported in column of Table 5. This data is also shown in Figure 2. The striking feature of this data is the very low degree of of overlap between the two groups: a Mann-Whitney U-test confirmed that the Petersen estimates reliably distinguish the two groups. (U=9.5, p<.01).

Table 5: Mean Peterson estimates based on the number of types in two tasks.

		Petersen estimate
Gp Advanced	**Mn**	160.37
	sd	*38.51*
Gp Intermediate	**Mn**	89.81
	sd	*31.3*

Figure 2 : Petersen estimates: SDs and outliers: two groups.



5: DISCUSSION

In this section, we will discuss some issues which arise out of the results reported in the previous section. Three important issues need to be highlighted. These are: a) the validity of the general approach, and b) whether the Petersen estimates give us any additional information which is not available in the raw word counts. The final section will consider a number of smaller issued raised by the data.

a) the general approach

In the introduction to this paper, we argued that we might be able to use methods developed for estimating animal population sizes as a way estimating the extent of vocabulary resources in L2 speakers. The data reported in section 4 suggests that this analogical extension of the species counting method has been partly successful, but not entirely so. The main finding is that the Petersen estimates generated from our raw data are clearly able to distinguish between the advanced and the intermediate groups, and that these estimates distinguish the groups rather better than the raw scores do. In all cases, the Petersen estimates suggest that the Ss' productive vocabulary is considerably higher than the actual counts we find in the raw data, and in this respect the method is clearly able to detect knowledge of vocabulary which is not immediately obvious in the raw data. However, as an estimate of overall vocabulary knowledge, the Petersen estimates are clearly not as helpful as we had hoped. The estimates suggest that our intermediate group has a productive vocabulary of about 90 words, and that our advanced group has a productive vocabulary of about 160 words. These estimates are clearly far too low to be interpreted at face value. We need to ask therefore, why the estimates have not produced more realistic figures.

With hindsight, it is obvious that Petersen estimates are very highly constrained by the number of types that are "trapped" by the data gathering method. The maximum value of the estimate is in fact determined by the product of the two data collection counts, M and N. Thus, if we collect 100 types for M, and 100 types for N, the maximum value of E is 100*100 = 10,000. In practice, this maximum would only be achievable if there was an overlap of 1 word type between the two data collections, and because of the repetitive nature of language, this is a highly unlikely occurrence. Even a very small degree of overlap between the two data collections would reduce our maximum value by a considerable amount. With only five words occurring in both texts, our estimate of the Ss' vocabulary size would fall to 2,000 words. With twenty words common to both texts, our estimate falls to 500 words. Our narrative description task actually elicited far fewer word types than this – for the advanced group, it generated just over 70 word types for each text, giving a maximum estimate value of about 4,900 words. However, the nature of the task meant that it was almost impossible to avoid using some of these words in both texts – *man, boy, stick, dog, throw, water*, as well as the obvious function words. For the advanced learners, about half of the word types found in Text 1 were also found in Text 2, giving a mean Petersen estimate of only 160 word types.

An alternative approach would be to exclude from our counts words which appear more than once in a text, on the grounds that these words are unavoidable components of the narrative, and do not really reflect the vocabulary items available to the subjects. This adjustment has the effect of reducing the values of M and N by about 50% - about half the words in a text typically occur only once. However, it also reduces the number of words which appear in both texts. This decreases the divisor in the Petersen formula, and accordingly increases the size of the Petersen estimate. For example, if we have two texts which each contain 100 words that occur once, and the number of words occuring in both texts is only 10, then the Petersen estimate estimate works out at

$$E = 100 * 100 / 10 = 10,000 / 10 = 1000$$

a figure which looks a lot more plausible than the estimates we reported earlier.

It seems that the choice of task here was more problematical than we realised. The narrative description task did not actually elicit much text, and the constraints of the narrative meant that there was a high probability that words elicited in Text 1 would also be elicited in Text 2. In terms of our animal species analogy, what we have here is a poor trapping device, one which tends to trap the same species twice, but leaves large numbers of other species out of account. Clearly, in future evaluations of this approach, we need to develop a test instrument that elicits longer texts, and is less likely to generate identical word types on both data collection occasions.

It seems to us that “word traps” of this sort need to take into account a number of factors which were missing from this exploratory study. Firstly, the elicitation instrument needs to be aware of the size of the productive vocabulary which we think our subjects have at their disposal. That is, if we think that we are dealing with a group of Subjects whose productive vocabulary is around 5,000 words, then we need to have an elicitation instrument which is capable of returning an estimate which is in this general ball-park. Secondly, we also need to take into account the fact that “word traps” which elicit continuous text will inevitably elicit words which appear in the two separate test events. Let us suppose that we could normally expect about 50% of the word types that appear in Text 1 to appear again in Text 2. In these circumstances, a word trap which elicits about 100 words of running text will typically produce a Petersen estimate of about 200 words – far too few to be a realistic estimate of a subject's productive vocabulary. On the other hand, a word trap which typically elicited more words, with a relatively small number of types that appear in two sequential data sets, might be capable of measuring much larger vocabularies. For example, a task that elicited 200 word types on each test occasion, with an overlap of only 10% of word types appearing in both data sets would, in principle be capable of producing reasonable estimates for a productive vocabulary of about 2,000 items. A task that elicited 250 words on each test occasion with only a 5% overlap on two test occasions might be capable of producing reasonable estimates for a productive vocabulary of around 5,000 words. We think that it might be possible to design a word trap of this sort using the methodology developed Fitzpatrick and Meara in their Lex30 test, and our guess is that a relatively small test of this sort might be capable of providing reasonable vocabulary size estimates over a wide range of L2 proficiency levels. Meara and Miralpeix's *Vocabulary Size Estimator* program, for example, suggests that intermediate level students typically have a productive vocabulary size of about 3500-6000 words. This range which could easily be assessed using a well-designed word trap based on a word association methodology, instead of the continuous text instrument used in this study.

b) what the Petersen estimates mean

It would be wrong, however, to give the impression that the Petersen estimates elicited in the present study are completely useless because the figures they generate are clearly not measuring the full extent of the productive vocabulary available to the Ss tested here. Our guess is that the estimates may still be providing us with useful information.

Firstly, it is possible that the Petersen estimates are telling us something about the productive vocabulary which is available to Ss **for this particular task**, and if this is correct, then the low level of the estimates might not actually be a serious problem. It would be relatively easy for us to collect data from groups of native speakers doing the same task, compute Petersen estimates for them, and then to compare the estimates we get for native speakers with the estimates our L2 speakers produce. For example, if we find that native speakers performing our narrative task typically generate Petersen estimates of, say, 350 words, with a standard deviation of 20 words, then we could report our L2 learner scores as a percentage of this native speaker score, or as a standardised score based on the native speaker mean and standard deviation. This looks like the beginnings of a methodology which would allow us to produce objective scores for the vocabulary used by L2 learners in productive tasks. The methodology might also enable us to assess the suitability of specific tasks used in vocabulary testing. For example, if the narrative task shown in Figure 1 turns out to generate very low productive vocabulary estimates when it is used with native speakers, we might want to conclude that it is not really appropriate as a tool for assessing L2 speakers' productive vocabulary size.

Secondly, it is possible that the Petersen estimates reported in section 3 may be good enough to act as an ordinal scale, even if they cannot be interpreted as absolute numbers. Clearly, the Ss in the advanced group have bigger productive vocabularies than the Ss in the intermediate group, and it is possible that the rankings produced by the Petersen estimates reflect the relative sizes of the Ss' vocabularies. It is also possible that there might be a fairly straightforward relationship between each S's Petersen estimate, and their actual productive vocabulary size if we could measure it. What we need to do here is to compare these results other estimates of productive vocabulary size, e.g. the estimates produced by Meara and Miralpeix's *Vocabulary Size Estimator*, or Malvern and Richards' (1997) *vocd* measure, and see whether there is a close correlation between these sets of measures. This work lies beyond the scope of a short paper of this sort.

c) other detailed points

A number of other minor points are worth discussing here.

Firstly, the significantly higher number of tokens and types produced by the Intermediate group on the second test is surprising. While the advanced group produce texts which look very homogeneous from the point of view of the number of word types they contain, the intermediate group seems to behave quite differently in this respect. All but one of the Ss in this group generated more word tokens in their second text than in their first texts, and all but one S produced a higher number of word types in text 2 - in some cases nearly double the number of word types. The advanced Ss are much more varied in this respect – about half the Ss show an increase in the number of word types from text 1 to text 2, while the other half show a reduction. This is an unexpected result which does not have an obvious explanation. We might have expected that performing the same task twice would have reduced the length of the narratives, and so reduced the number of word types contained in the second text, but this does not appear to be the case for the intermediate learners. For tokens, the standard deviation for the learners in Test 1 is much smaller than the other three standard deviations, but again it is difficult to work out what this might mean. For types, we have a very similar pattern of results. There is clearly a need for more work on repeated tasks of this sort if we are to work out whether this pattern of performance is a reliable feature of intermediate level learners or not, and whether the sort of random variation (and large sds) that we get with the more advanced Ss is typical of how Ss behave

Secondly, the groups appear to differ in the number of words that appear in both texts. (t=2.71 p=0.13), with the advanced group having a much larger number of repeated words than the intermediate group. Again, this is not necessarily what we would have expected. We might have expected that the number of repeated words in the two groups would have been very similar – basically we might have expected that both sets of texts would repeat function words and a small number of unavoidable words required by the narrative, but the degree of repetition found here seems to go beyond this. Ironically, this tendency has the effect of lowering the Petersen estimates for this group. All other things being equal, the bigger the number of repeated words, then the lower the resulting Petersen estimate will be, and this makes the significant difference between the groups even more striking than it appears at first sight.
Again, what we need here is some further research into how likely it is for word types to reappear in repeated tasks, and how this interacts with vocabulary size.

Thirdly, we need to ask whether the Petersen estimates actually tell us anything that we could not already work out from the raw data presented in Table 2.

All the main variables distinguish between the groups: number of tokens in text 1 (t=5.50, p<.001); number of tokens in text 2 (t=2.94, p<.001); number of types in text 1 (t=5.38,p<.001); number of types in text 2 (t=3.38,p<.001);peterson estimate (t=5.30, p<.001). The pattern of results here is again slightly odd, with the text1 scores and the Peterson estimates returning the best values. The values for text 2 also distingush the groups, but are not so striking.

Table 6 shows the correlations between the Peterson estimates and the the other variables .

The data show that the correlations between the raw token counts and the Petersen estimates are generally fairly good, 0.758 for text 1, 0.671 for text 2, and rising to .801 for the two texts combined. For types the correlations are slightly higher - though given

Table6: correlations between the Peterson Estimate and the other variables

variable	**Tokens T1**	**Tokens T2**	**Tokens T1+T2**	**Types T1**	**Types T2**
correlation all Ss	.758	.671	.801	.823	.861
intermedts	.273	.524	.477	.645	.897
advanced	.506	.476	.611	.601	.726

that the Peterson estimates depend very heavily on the Type data, this is perhaps not surprising. When we look at the groups separately, we find that the correlations are generally more consistent for the advanced group, with one very striking correlation between Types on text 2 and the Peterson estimates. Again this points to something slightly odd about the way the intermediate learners approach the second story-telling task.

These findings suggest that the Petersen Estimates may indeed be tapping into some interesting features of vocabulary use by L2 speakers, but it is not staightforward to work out exactly what these features are without collecting a lot more data.

6: CONCLUSION

This paper has looked at the use of Petersen Estimates as a way of assessing how much productive vocabulary L2 learners have at their disposal. The data suggests that there might be ways of constructing effective "word traps" which can be used to make realistic estimates of productive vocabulary size, and that taken together with other estimates these tools might be able to generate plausible estimates of productive vocabulary size in L2 speakers. Standard essays do not appear to be a good way of collecting the relevant data – it is difficult to collect long essays which generate large numbers of different words, and unavoidable repetition of function words and key vocabulary items means that the Petersen estimates are a lot smaller than we would expect. Nevertheless, the general approach seems to hold out considerable possibilities, and we think that it might be possible to develop alternative trapping methods based on word association techniques. We will be exploring this methodology in a future paper.

REFERENCES

Laufer, B
The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics* 19,2(1998), 255-271.

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

Lincoln, FC
Calculating waterfowl abundance on the basis of banding returns. *Cir. US Department of Agriculture* 118(1930) 1-4.

Malvern, DD and BJ Richards
A new measure of lexical diversity. In: **A Ryan and A Wray** (eds.) *Evolving Models of Language*. Clevedon: Multilingual Matters. 1997: 58-71.

Meara, PM and T Fitzpatrick
Lex30: an improved method of assessing productive vocabulary in an L2. *System,* 28,1(2000), 19-30.

Nation, ISP
Teaching and learning vocabulary. Rowley, Mass.: Newbury House. 1990.

Meara, PM and H Bell
P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16,3(2001), 5-24.

Melka Teichroew, FJ
Receptive versus productive vocabulary: a survey. *Interlanguage Studies Bulletin* 6,2(1982), 5-33.

Melka Teichroew, FJ
Les notions de réception et de production dans le domaine lexicale et sémantique. [Receptive and Productive vocabulary.] Berne: Peter Lang.

Melka, F
Receptive vs productive aspects of vocabulary. In: **N Schmitt and M McCarthy** (eds.) *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press. 1997.

Meara PM and I Miralpeix
Vocabulary size estimator. Swansea: Lognostics. 2007.

Petersen, CGJ
The yearly immigration of young place into the Linsfiord from the German Sea. *Rep. Dan. Biol. Stn* 6(1896), 5-84.

Seber , GAF
The estimation of animal abundance and related parameters. London: Edward Arnold. 1982.

Seber, GAF
A review of estimating animal abundance. *Biometrics* 42(1986), 267-292.

Slagter, P
Un nivel umbral. Strasburg: Council of Europe. 1979.

Zipf, GK
Psycho-Biology of Language. New York: Houghton-Mifflin. 1935.

First posted: August 1st 2008.