*Reading in a Foreign Language* April 2010, Volume 22, No. 1
ISSN 1539-0578 pp. 222–236

# Words as species: An alternative approach to estimating productive vocabulary size

Paul M. Meara
Swansea University
United Kingdom

Juan Carlos Olmos Alcoy
University of Dundee
United Kingdom

## Abstract

This paper addresses the issue of how we might be able to assess productive vocabulary size in second language learners. It discusses some previous attempts to develop measures of this sort, and argues that a fresh approach is needed in order to overcome some persistent problems that dog research in this area. The paper argues that there might be some similarities between assessing productive vocabularies—where many of the words known by learners do not actually appear in the material we can extract them from—and counting animals in the natural environment. If this is so, then there might be a case for adapting the capture-recapture methods developed by ecologists to measure animal populations. The paper reports a preliminary attempt to develop this analogy.

***Keywords***: productive vocabulary, capture, recapture, word counts, ecological models

Paul Nation's (1990) Vocabulary Levels Test has perhaps been the single most important development in vocabulary acquisition research in the last 20 years. The test provides a rough estimate of a learner's receptive vocabulary size in the form of a vocabulary profile. Simple to use, and easy to understand, it has been widely adopted by researchers around the world, and has rapidly become the *de facto* standard vocabulary size test. The vocabulary size estimates that it produces appear to be remarkably reliable and robust. This has led to the Vocabulary Levels Test being used in a very large number of empirical studies where vocabulary size is a critical variable, and particularly in studies that have examined the relationship between vocabulary size and reading ability in second language (L2) learners. Inevitably, however, the development of a standard assessment tool of this sort opens up other areas of research, and the Vocabulary Levels Test is no exception to this generalisation. The availability of a reliable measure of receptive vocabulary size leads to some very interesting questions about the relationship between receptive vocabulary and active productive vocabulary. This issue is one extensively addressed in Nation's work.

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 that need to be acquired productively and other vocabulary items that only need to learned for recognition purposes. Despite the fact that many researchers have written about this topic at a theoretical level (Corson, 1983, 1995; Laufer, 1998; Melka, 1997; Melka Teichroew, 1982, 1989), the idea of productive vocabulary remains a fundamentally elusive one. The main reason for this is that it has proved surprisingly difficult to develop simple and elegant tests of productive vocabulary size that have any degree of face validity, and this makes it difficult to answer, with confidence, questions such as *How are receptive and productive vocabulary related? Do receptive and productive vocabulary grow at the same rate? Are there thresholds in the development of a passive vocabulary?* Not surprisingly, perhaps, given the widespread use of the Nation's Vocabulary Levels Test to assess receptive vocabulary, the approach most widely used in the recent research literature that investigates productive vocabulary in L2 learners is an adaptation of the original Vocabulary Levels Test usually known as the *Productive Levels Test* (Laufer & Nation, 1999). Laufer has used these two tests in combination to make some very interesting theoretical claims about the relationship between 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 are much more difficult to interpret than the data provided by the original Vocabulary Levels Test, and in our view it is worthwhile looking at alternative approaches to estimating productive vocabulary size. This is not to denigrate the usefulness of the Productive Levels Test approach, of course, but rather because we think that *productive vocabulary* may be a more complicated notion than it appears to be at first sight, one that would benefit from being examined from a number of different and perhaps unconventional points of view.

In our previous research, we have developed three main ideas, which we think might allow us to "triangulate" the idea of productive vocabulary size. For obvious reasons, most traditional studies of productive vocabulary require learners to produce short texts for evaluation, but this material is difficult to collect, particularly when you are dealing with low level learners who are reluctant to produce extended texts. Our first solution to this problem was to move away from using written texts as the raw data for research on productive vocabulary size. We (Meara & 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 the true indicators of a large productive vocabulary. We tried to get round this problem by getting learners to generate "texts" derived from a set of word association tests called *Lex30*. These data typically consisted of relatively infrequent L2 words that could be profiled using standard vocabulary assessment tools such as Range (Heatley, Nation, & Coxhead, 2002), and we argued that these profiles provided a better picture of the scope of a testee's productive vocabulary than other, more traditional test types did. Unfortunately, although the test scores tended to correlate with tests of receptive vocabulary size, it was not obvious how the profiles provided by the Lex30 test could be converted into proper estimates of productive vocabulary size.

In our second approach to estimating productive vocabulary (Meara & Bell, 2001), we returned to using texts generated by L2 writers, and attempted to develop an "extrinsic" measure of vocabulary richness. This paper analysed sets of short texts produced by L2 learners, and for

each text generated a curve that described the incidence of "unusual" words in short segments of text. We 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 & Miralpeix, 2007) we 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 that closely matched the actual data produced by the learner. This general approach proved to be solid enough to distinguish between advanced and less advanced learners. More importantly, however, this approach actually allows us to quantify the productive vocabulary that seems to be behind a particular text. For example, it allows us to tentatively make statements like "the text in Example 1 implies a productive vocabulary of around 6,400 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.

*Example 1.* V-Size estimates that the following text was generated by a speaker with a productive vocabulary of at least 6,400 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.

Obviously, it would be very useful to have a tool that would allow us to estimate a learner's productive vocabulary size with some degree of confidence. For this reason, 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 that 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 take further the ideas raised in Laufer's (1998) paper. This paper sketches an approach that is rather different from the approaches we have developed in our

previous work, but one that we feel is very much in the spirit of Paul Nation's *thinking outside the box* approach to vocabulary testing.

## 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 participants. If we were dealing with learners with very small vocabularies, then it might be possible to devise a set of tests that assessed whether our learners could produce each of the words in a short list of target words that we are interested in. In practice, however, this only works where we are dealing with very small vocabularies. In real testing situations, it is logistically impractical to test the entire vocabulary of a learner who has more than a very elementary vocabulary. In this paper, for example, we are interested in learners of Spanish. *Threshold Level Spanish* (Slagter, 1979) comprises a lexicon of around 1,500 words, which gives learners only a very limited level of competence in Spanish. Testing vocabulary exhaustively at this level is difficult, though it is just about feasible with very co-operative participants. Testing the vocabulary of more advanced participants becomes increasingly difficult as their vocabulary grows. Consequently, if we want to test the vocabularies of even moderately 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 that 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 that 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 straightforward: The animals are not co-operative and do not line up in a way that allows us to number them reliably. This makes it notoriously difficult to make good estimates of animal populations, a problem that can have serious consequences if we are trying to manage the population and control the number of animals that a particular environment can provide for, or as in the case of the cockroaches, we have to eliminate them altogether.

Ecologists have developed a number of methods that allow them to resolve 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 fish in a river. We could identify a suitable stretch of river to investigate, and then distribute traps that will catch the fish without harming them. We leave the traps out for a set time, overnight, for instance, and count the number of fish that we have trapped. We then mark these animals in a way that will allow us to identify them, before releasing them back into the

wild. The next night, we carry out the same counting exercise, enumerating the fish trapped overnight. This gives us three numbers: We have N, the number of fish captured on Day 1; M, the number of fish captured on Day 2; and X, the number of fish that were captured on both occasions. Petersen argued that it was possible to extrapolate from these figures to the total number of fish in the stretch of river. Petersen's estimate is calculated as follows:

$$E = (N * M) / X$$

That is, Petersen's estimate of the size of the fish population is the product of the two separate counts divided by the number of fish counted on both occasions. A simple example will make this idea more concrete. Suppose that on Day 1 we count 100 fish in a 10 mile stretch of river, and we mark them all. On Day 2, we find 60 fish, 20 of which were also noted on Day 1. Petersen's estimate of the number of fish inhabiting the stretch of river would be

$$E = (100 * 60) / 20 = 6{,}000 / 20 = 300$$

If the river is actually 100 miles long, with similar conditions throughout, then our 10 mile stretch represents a 10% sample of the whole river, so we could extrapolate that there are about 3,000 fish in the entire length of the river.

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 fish was a reasonable one, one that gave us a good chance of capturing the fish we want to count, and that the one mile stretch we have selected represents in some way the entire river. 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 fish needs to be constant from Day 1 to Day 2—if half our fish were killed by otters, or died from poisoning overnight, then Petersen's model would simply not apply. Finally, we are assuming that the data collection on Day 2 is "equivalent" to the data collection on Day 1, 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 fish in the river, even though we are not able to count every single one of them, and even though we only sampled a part of the entire river.

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, 1986) provided 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 fish 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 that 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 animal species we find in our stretch of river, rather than the number of fish. Suppose that our first data collection event delivers 10 different types of animal, and we make a record of these 10 types. If our second data collection delivers 12 types of animal, of which 8 were previously recorded, then Petersen's estimate of the number of species inhabiting the river 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 fish *types*, rather than the number of different fish *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 that 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 the final section of this paper.

## Methodology

### *Participants*

Twenty-four participants took part in this study. All of them were learning Spanish at the University of Dundee. Eleven of the participants were at a low intermediate level, while the remaining 13 participants were considered by their teacher to be "advanced." These participants were all native English speakers. We acknowledge that these numbers are very small. We also acknowledge that there are some important differences between Spanish and other languages, which may have affected the results.

### *Data Collection*

The 24 participants were asked to write a description of a cartoon story. The story consisted of six pictures. In the first picture, a man and a boy are playing with a dog beside the sea. The boy throws a stick into the sea for the dog to fetch. The second picture shows this game being observed by a smartly dressed man with an umbrella. In the third picture, this man approaches the dog and shows it his umbrella. The fourth picture shows the smart man throwing his umbrella into the sea. Unfortunately, the dog ignores this. In the fifth picture, the man, the boy, and the dog abandon the smart man, leaving his umbrella floating on the water. The final picture shows the smart man removing his clothes, presumably so that he can swim out to sea and rescue his lost umbrella.

Participants 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 the participants were asked to write a second description of the same cartoon story. In both data collection events, participants write their stories by hand. The hand-written stories were then collected and transcribed into machine readable format for further analysis. Example 2 illustrates the kind of material that was generated by this task. The use of a single time-limited task is an analogue of the method used to count fish in the river. We are not looking for a task that will elicit every single word a participant knows: Rather we are trying to devise a word trap that will capture enough words for us to make a reasonable estimate of the participant's vocabulary.

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 programme that 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 that 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*, and *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.

*Example 2.* Below is a sample text elicited by Figure 1.

> Hay un hombre y un niño cerca de un río y el hombre está mirando el niño, el niño está jugando con el perro y se tira un ayuda de andar de madera en el río.
>
> 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 está 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 están mirando a el hombre formal.
>
> El hombre formal empieza a tirar su ayuda de andar de madera en el río, con gran fuerza, se usa todo su cuerpo para tirar y el madera va muy, muy lejos en el río. El hombre original, el niño y el perro estan mirando, sin movimiento, a el hombre formal.
> Ahora el ayuda de andar de madera está en el río, 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 está muy discontenta, su madera está lejos y en el río. El hombre formal pregunta a el perro, el hombre y el niño para que queden y el perro trae el madera del río.
>
> Ahora el hombre formal está solo y está mirando el ayuda de andar de madera pero al

mismo tiempo está sacando todo su ropa para que nade a su madera. Su sombrero, zapatos, chaqueta y camiseta están en el suelo y ahora mismo el hombre formal está sacando sus pantalones.

## Results

Table 1 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, $t(10) = 3.37$, $p < .01$, though the Group × Test Time interaction is not significant.

These data are 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 the next section of this paper.

Table 1. *Mean number of word tokens in two narrative description tasks*

| Group | T1 narrative | T2 narrative | Combined |
|---|---|---|---|
| Advanced | | | |
| Mean | 190.23 | 199.15 | 389.38 |
| *SD* | 48.72 | 63.65 | 59.81 |
| Intermediate | | | |
| Mean | 99.19 | 133.63 | 232.81 |
| *SD* | 27.16 | 40.28 | 89.94 |

Table 2 shows a more complex data set that records for each participant the number of different *word types* they produced in each of the data collections, along with the number of word types that occurred in both narratives.

Table 2. *Mean number of word types in two narrative description tasks*

| Group | T1 narrative | T2 narrative | Combined |
|---|---|---|---|
| Advanced | | | |
| Mean | 72.91 | 73.73 | 33.55 |
| *SD* | 17.00 | 19.09 | 9.11 |
| Intermediate | | | |
| Mean | 43.36 | 52.36 | 25.82 |
| *SD* | 8.89 | 15.09 | 6.91 |

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 significant test effect, or any significant Group × Test interaction.

For each participant, the raw number of types was plugged into the Petersen estimate formula, and the estimates generated in this way are reported in Table 3. The striking feature of these data is the very low degree 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 3. *Mean Petersen estimates based on the number of types in two tasks*

| Group | Petersen estimate |
|---|---|
| Advanced | |
| Mean | 160.37 |
| *SD* | 38.51 |
| Intermediate | |
| Mean | 93.81 |
| *SD* | 31.30 |

## Discussion

In this section, we will discuss some issues that arise out of the results reported in the previous section. Two 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 that is not available in the raw word counts. The final section will consider a number of smaller issues raised by the data.

### *The General Approach*

In the introduction to this paper, we speculated 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 suggest 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 token counts and raw type counts do. In all cases, the Petersen estimates suggest that the participants' 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 that 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. The figures suggest that the vocabulary of the advanced participants is nearly twice that of the intermediate

participants, which seems plausible. However, the absolute figures are just ridiculously low, and clearly they cannot 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 participants' 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 that 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 participants. 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 that 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 occurring in both texts is only 10, then the Petersen estimate works out at

$$E = 100 * 100 / 10 = 10{,}000 / 10 = 1{,}000$$

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

It seems then 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 that 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 that were missing from this initial exploratory study. Firstly, the elicitation instrument needs to be aware of the size of the productive vocabulary that we think our participants have at their disposal. That is, if we think that we are dealing with a group of participants whose productive

vocabulary is around 5,000 words, then we need to have an elicitation instrument that is capable of returning an estimate that is in this general ballpark. Secondly, we also need to take into account the fact that word traps that elicit continuous text will inevitably elicit words that 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 that 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 an advanced participant's productive vocabulary. On the other hand, a word trap that 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 by 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 3,500–6,000 words. This range 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.

*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 participants 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 participants *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 that 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 participants in the advanced group have bigger productive vocabularies than the participants in the intermediate group, and it is possible that the rankings produced by the Petersen estimates might reflect the relative sizes of the participants' vocabularies. It is also possible that there might be a fairly straightforward relationship between each participant'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, for example, 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 exploratory paper of this sort.

*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 produces texts that 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 participants in this group generated more word tokens in their second text than in their first texts, and all but one participant produced a higher number of word types in text 2—in some cases nearly double the number of word types. The advanced participants are much more varied in this respect—about half the participants 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 standard deviations) that we get with the more advanced participants is typical of how participants normally behave.

Secondly, the groups appear to differ in the number of words that appear in both texts, $t = 2.71$, $p < .05$, 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 anticipated. We might expect that the number of repeated words in the two groups would have been very similar—basically we might expect 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, 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 tendency 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 1. 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, $t = 3.38$, $p < .001$; and Petersen estimate, $t = 5.30$, $p < .001$. The pattern of results here is again slightly odd, with the text 1 scores and the Petersen estimates returning the best values. The values for text 2 also distinguish the groups, but are not so striking.

Table 4 shows the correlations between the Petersen estimates and the other variables.

Table 4. *Correlations* (*r*) *between the Petersen estimates and other variables*

| Participant | Token | | | Type | |
|---|---|---|---|---|---|
| | T1 | T2 | All | T1 | T2 |
| All | 0.758 | 0.671 | 0.800 | 0.823 | 0.861 |
| Intermediate | 0.273 | 0.524 | 0.477 | 0.645 | 0.897 |
| Advanced | 0.506 | 0.476 | 0.611 | 0.601 | 0.726 |

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 .800 for the two texts combined. For Types the correlations are slightly higher—though given that the Petersen 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 Petersen 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 straightforward to work out exactly what these features are without collecting a lot more data.

## 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 suggest 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.

At first sight, it might look as though productive vocabulary does not have much to say about reading in a foreign language. However, a large body of research suggests that a significant component of reading skill is the ability to predict the content of a text, and to anticipate the occurrence of words a text contains. It seems highly likely that there will be some sort of relationship between the *predictive vocabulary* skills needed for reading and *productive vocabulary* in general, perhaps even a stronger one than we find between reading and receptive vocabulary size. We hope that the work reported here, despite its very obvious limitations, might turn out to be a small step in the direction of teasing out the nature of this relationship.

## References


Corson, D. J. (1983). The Corson measure of passive vocabulary. *Language and Speech*, *26*, 3–20.

Corson, D. J. (1995). *Using English words*. Dordrecht: Kluwer Academic.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

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

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

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

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, England: Multilingual Matters.

Meara, P. M., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*, 5–24.

Meara, P. M., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*, 19–30.

Meara, P. M., & Miralpeix, I. (2007). *Vocabulary size estimator*. Swansea: Lognostics.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Melka Teichroew, F. J. (1982). Receptive versus productive vocabulary: A survey. *Interlanguage Studies Bulletin*, *6*, 5–33.

Melka Teichroew, F. J. (1989). *Les notions de réception et de production dans le domaine lexicale et sémantique*. Berne: Peter Lang.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

Petersen, C. G. J. (1896). The yearly immigration of young plaice into the Linsfiord from the German Sea. *Rep. Dan. Biol. Stn*, *6*, 5–84.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*. London: Edward Arnold.

Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics*, *42*, 267–292.
Slagter, P. (1979). *Un nivel umbral*. Strasburg: Council of Europe.
Zipf, G. K. (1935). *Psycho-biology of language*. New York: Houghton-Mifflin.

## About the Authors

Paul M. Meara is Professor of Applied Linguistics at Swansea University where, until his recent retirement, he was in charge of the Vocabulary Acquisition Research Group.
E-mail: P.M.Meara@swansea.ac.uk

Jaun Carlos Olmos Alcoy completed his PhD at Swansea in 2009. He teaches Spanish at the University of Dundee in Scotland. E-mail: j.c.z.olmosalcoy@dundee.ac.uk