DOI: 10.7763/IPEDR. 2013. V68. 5

The Schnabel Method: An Ecological Approach to Productive Vocabulary Size Estimation

Juan Carlos Olmos Alcoy[+]

Mahidol University, Mahidol University International College, Salaya, Thailand

Abstract. This study addresses the issue of how we might be able to quantify productive vocabulary size in two groups of university students learning Spanish as L2. One group has an intermediate level of Spanish; the other is advanced. 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. Hence, an approach used to quantify vocabulary density in this paper –the Schnabel method- has been extrapolated from population ecology to estimate lexical density.

Keywords: Productive Vocabulary, Capture-recapture, Schnabel Method, Ecological Models

1. Introduction

In recent years, there has been an increasing interest in quantifying productive vocabulary in L2 research. However, there is a major practical problem associated with estimation of productive vocabulary size: we simply cannot ask L2 students to write all the words they know. Inevitably, researchers have to resort to extracting samples of the productive vocabulary students know and, then, find a way to estimate overall size from these samples. This problem is not one which is restricted to Applied Linguistics, however. Exactly the same issues arise in ecological studies of animal abundance, where researchers have developed a number of statistical methods which allow them to estimate the size of animal populations from partial samples (Seber, 1973; Begon, 1979; Krebbs, 1999; Pollock, 2002).

2. Ecological Approach to Productive Vocabulary Size Estimation.

2.1. Basic Model and Assumption

In ecology there are a few well-established mathematical models used to estimate the total size of particular target populations. This is not carried out by trying to count every single member of the population under scrutiny. This approach has been tried before and it has proved to be unreliable (Begon, 1979). Instead, there are other techniques which rely on samples from the target population. These samples provide the springboard for researchers to make estimates about the size of the whole group. These ecological techniques that rely on samples are called *capture-recapture* or *mark-recapture.*

Let us see, in practical terms, how the basic *capture-recapture* technique works. It is called the *Petersen estimate* and we only need two samples to use it. We are interested in finding out how many animals of a certain species there are in a particular area. This number is an unknown figure and it is often represented as ***N***. One day a number, that is, a sample of these animals is captured, marked and then released into the same area where they were captured. This number will be represented as ***C***. The following day another sample of the same animal species is captured again; this figure will be called ***M***. Some of them will have a mark because they were also captured on the first day. This number will be represented as ***R***. So, if we compile all the data, we have:

N = total number of animals (unknown)
C = number of animals captured on the first day
M = number of animals captured on the second day
R = number of animals captured both days

[+] *E-mail*: juancarlos.olm@mahidol.ac.th

The assumption underlying *capture-recapture* techniques is that the proportion of marked individuals recaptured in the second sample represents the proportion of marked individuals in the population as a whole. In mathematical terms:

$$\frac{N}{C} = \frac{M}{\mathrm{R}}$$

This is how the *Petersen estimate* method is applied in statistical ecology. Now, we are in a position where we can try to adapt it to vocabulary estimation. Let us imagine we have given a student of English as L2 a productive task (**Task 1**) and s/he produces, say, 948 words. This number is only a sample of all the words this particular student knows productively. The data gathered so far can be represented as ***C*** = 948 and ***N*** = unknown. The next step is as follows: a few days later, the same productive task (**Task 2**) is given to the same student who completed Task 1. On this occasion, the student produces, say, 421 words. There is also another useful piece of information we can extract now by comparing the productions of **Task 1** and **Task 2**: the number of words that appear in both tasks. Let us assume this number is 167. The extra data gathered can be represented as ***M*** = 421 and ***R*** = 167. Now, if we make the pertinent extrapolations, each letter means:

N = estimate of the total number of words the student knows productively.
C = number of words elicited by the student in **Task 1.**
M = number of words the student elicited in **Task 2.**
R = number of words in **Task 2** also present in **Task 1.**

$$\frac{N}{C} = \frac{M}{\mathrm{R}}$$

$$\frac{N}{948} = \frac{421}{167} = \frac{948 * 421}{167} = 2{,}389.86$$

We would conclude, then, that this particular student has an overall productive vocabulary size of approximately 2,390 words in English as L2.

2.2. Conditions of the *Capture-recapture* Models.

There are three conditions that the target population has to meet in we want to achieve reliable results. These are **a)** the target population size needs to be constant, **b)** the sample is random, and **c)** all animals have the same chance of being captured. If we extrapolate these three conditions into our lexical estimation context, we obtain **a)** the vocabulary size students know remains constant between **Task 1** and **Task 2** **b)** the sample of vocabulary elicited by the student(s) is random, and **c)** all words have the same chance of being elicited.

2.3. The Schnabel method.

The *Petersen estimate* relies only on two sampling occasions and it is easy to apply. However, it has a drawback: it tends to overestimate the population size (Seber, 1973; Begon, 1979; Krebbs, 1999). To reach more accurate estimates, ecologists often use techniques needing multiple marks and recaptures. One of these techniques is the *Schnabel method*, which is not markedly dissimilar to the *Petersen estimate*. The main difference is that it allows for more than 2 capture-recapture encounters. It can already be applied with a minimum of 3 captures and there is no limit as to the maximum number of recaptures required. The *Schnabel method* only distinguishes two types of individuals: *marked* = caught in one or more prior samples; and *unmarked* = never caught before (Krebbs, 1999; p. 35). Algebraically, the formula is:

$$N = \frac{\sum_{i=1}^{m} M_i C_i}{\sum_{i=1}^{m} R_i}$$

where *Mi* = the total number of previously marked animals at time *i, Ci* = the number caught at time *i,* and *Ri* = the number of marked animals caught at time *i.* Below we have an example of how this data can be tabulated:

Table 1: Data from a population of Cricket Frogs *Acris gryllus* in Louisiana, USA, sampled over 5 successive days (Adapted from Sutherland, 2006; p. 26).

	Number of animals caught	**Number of recaptures**	**Number of new animals caught**	**Total number of tagged animals**[1]
1st capture	32	0	32	0
2nd capture	54	18	36	32
3rd capture	37	31	6	68
4th capture	60	47	13	74
5th capture	41	36	5	87
Total	224	132	92	261

The formula needed to estimate the size N of the population is really an extension of the formula used in the *Petersen estimate*. We will refer to the data above in order to illustrate the *Schnabel method.* In each row we multiply the number of animals caught (first column) by the total number of animals tagged (fourth column). Given the data above, we can do this operation 5 times, one per capture event. The results of all five multiplications are then added up. Finally the figure given by this addition is divided by the total number of recaptures (second column; last figure). This process is exemplified below:

$$\text{N} = \frac{(54*32)+(37*68)+(60*74)+(41*87)}{132} = \frac{12{,}251}{132} = 92.81$$

It is easy to see how this model can be adapted to estimate lexical size. Let us imagine a student performs the same productive task 3 or more times. The data provided by all these tasks can then be classified as follows: **a)** number of words used every time the task is completed, **b)** number of words used in previous tasks, **c)** number of new words used in every task, and **d)** total number of different words used. We are now in a position to apply the *Schnabel method* in a vocabulary size estimation context.

3. Experiment.

3.1. Subjects

A group of 43 students from (low) intermediate to advanced levels of proficiency participate. The students come from a variety of European countries.

3.2. Methodology

In order to maximise randomness we use the Roman alphabet to stimulate lexical productions. The alphabet is presented in a column. The alphabetical list is followed by 5 other columns. At the top of the page, students can read the following instructions: *"You have the Spanish alphabet below followed by 5 columns. You are required to write down Spanish words beginning with the letter on the left: start with* ***Columna 1*** *all the way down (i.e: Amigo, Bueno...). Once you have completed the first column, move on to the second. Continue in this fashion till you have completed all 5 columns. If at some point you cannot think of a word starting with a particular letter, don't stop; just move on to the next letter. You have a maximum of 30 minutes to complete the 5 columns."*

The task is completed 4 times over a period of three weeks. It is emphasised that: **a)** they should write the first word that comes to mind when they read each letter, and **b)** there are no right or wrong answers.

3.3. Counting words

The following criteria are applied when counting words:

[1] Note this is the number of caught animals **before** the capture event takes place.

- Instances of nouns and adjectives used more than once but with different gender or number markers (i.e: *bajo*, *baja*, *bajos*, *bajas*) are all counted as 1 word.
- Different instances of the same verb (i.e: *estoy*, *estaba*, *estuve*) are also counted as 1 entry when scoring is done.
- Proper nouns, numbers, abbreviations, function words and items which are not recognised as an existing Spanish word are ignored when counting is done.
- Minor spelling mistakes are corrected and taken into account.
- When polysemic items appear, the student is always given credit. If a student produces, say, *llamar* and *llama*. Since *llama* has also another meaning, they are counted as 2 words.

3.4. Results

After the students complete the task for the fourth time, words are counted and the *Schnabel method* is applied. Based on my knowledge of the students' level of proficiency, they are now divided into two ability groups: intermediate and advanced. In the table below we can see the mean results for each one of the groups:

Table 2: Mean results per group

		Overall number of words used	Total number of repetitions	Schnabel estimate
Int.	Mean	192.86	85.45	384.06
	Standard deviation	15.16	15.99	62.70
Adv.	Mean	239.47	57.19	805.79
	Standard deviation	22.80	19.82	313.24

A univariate analysis of variance is carried out between both groups. It shows a significant difference between both groups (F (1, 41) = 38.31, p < 0.01). This confirms that the advanced group knows significantly more productive words than the intermediate group. Results in both groups suggest that none of the three conditions is significantly violated.

3.5. Discussion

The results reported above raise a number of important issues that need to be addressed. The issues are a) the reliability of the *Schnabel method*, and b) to what extent the three conditions have been met.

Reliability of the Schnabel method: The results reported above imply that this approach is partially successful. Furthermore, most N estimates are (remarkably) higher than previous experiments carried out in the same area (Olmos Alcoy, 2009). This suggests we may be another step closer to achieving more realistic figures. The mean N estimate for the intermediate group is about 384 words, and for the advanced group is about 805 words. Within the advanced group, five students obtained more than 1000 words in the N estimate. This is a considerable improvement when compared to all the previous estimations carried out in this field. If we now apply the *Petersen estimate* (using the number of words elicited in the first 2 tasks) to the Schnabel method data, we can see how the two estimates compare. The two tables below provide a summary of the results:

Table 3: Summary of results for intermediate group

	Petersen estimate	**Schnabel method**
N	254.30	384.06
Standard deviation	45.43	62.70

As we can see, the *Schnabel method* gives us the highest estimates. This implies that this approach detects a significant amount of lexical knowledge that the *Peterson estimate* ignores. Still, if we apply an ANOVA to the *Peterson estimate*, we find that it can reliably discriminate between both groups (Peterson estimate: F (1, 41) = 22.95, p < 0.01).

Table 4: Summary of results for advanced group

	Petersen estimate	Schnabel method
N	441.67	805.79
Standard deviation	236.56	313.24

The standard deviations are rather problematic, especially with the advanced group (= 313.24). This may be due to two factors: **a)** the estimate of one of the students is unusually high (1888.52 words), and **b)** two students have a very proficient command of Spanish because they have lived in Spain for a few years. If we remove these students' data and re-calculate estimates, we obtain N= 729.43; *sd*= 187.92. The N estimate is still quite high and the *sd*, though, greatly reduced, remains quite high as well. This may suggest that there is a wide range of lexical knowledge within the L2 advance group.

To what extent the 3 conditions have been met: We will assess now to what extent each group is likely to have met the three conditions.

Condition 1: the number of words students knows between the completion of all 4 tasks remains constant. This condition is very likely to hold for both groups because all tasks were completed in a short period of time (20 days).

Condition 2: words are randomly produced. The amount of repetitions produced by each group gives us some indication of how random responses are. It is theorised that the fewer repetitions occur, the more likely responses are randomly elicited. Clearly, the intermediates re-use more items than the advanced group. It proves that the advanced students have greater lexical knowledge at their disposal; however, we should not stop here. The task was designed to maximise random responses. Some evidence of this was found after task completion. Students were interviewed about their responses; it was reported that most words were simultaneously produced on the spur of the moment for no apparent reason. Some students also pointed out that they had rarely, or never, produced some of the words before. This feedback suggests that to some extent random elicitation was achieved. Strictly speaking, we cannot say that these words show evidence of students' lexical proficiency because we do not know whether the students can produce them adequately in context or not. All we can say is that, on some level, these words are part of the students' productive vocabulary.

Condition 3: all words are equally eligible for elicitation. We know, however, this is not the case: most words have a different probability of being produced. Condition 3 is very likely to have been violated to some extent despite the fact the task was completely decontextualised. A regression plot was applied to both groups and, interestingly, the intermediates are rather close to giving a significant graph, unlike the advanced group. This may suggest that, in a task that purports to extract random samples of vocabulary, the responses of less proficient students are less affected by word frequencies. The less linear regression plot given by the advanced group may be seen, then, as an indicator of the students' greater use of words from different frequency bands.

4. Conclusion

In this paper we have seen how the *Schnabel method* can be used to estimate productive vocabulary size. We have also used the alphabet as stimulus to optimize random lexical elicitation. Results are encouraging to a degree because the N estimates are much higher than those of previous experiments (Olmos Alcoy, 2009). Future experiments should explore other approaches and methodologies in order to further enhance our overall N estimates.

5. Acknowledgements

I am very grateful to my PhD supervisor Prof. Paul Meara for his invaluable comments.

6. References

[1] Begon, M., 1979. *Investigating Animal Abundance: capture-recapture for biologists,* London, Edward Arnold

(Publishers) Limited

[2] Krebbs, J. C., 1999. *Ecological Methodology.* Addison-Welsey Educational Publishers, Inc.

[3] Olmos Alcoy, J. C. 2009. *Estimating productive vocabulary using models extrapolated from ecology.* PhD thesis. Swansea University.

[4] Pollock, K. H., 2002. Capture-Recapture Models. In Raftery, A. E., Tanner, M. A. and Wells, M. T., eds *Monographs on Statistics and Applied Probability 93.* United States of America: CRC Press LLC and American Statistical Association.

[5] Seber, G. A. F., 1973. The estimation of animal abundance and related parameters. Bristol: J. W. Arrowsmith Ltd.

[6] Sutherland, W. J. 2006. *Ecological census techniques.* Cambridge: Press syndicate of the University of Cambridge.