# ECOLOGICAL APPROACH TO VOCABULARY DENSITY ESTIMATION

Dr. Juan Carlos Olmos Alcoy*

## Abstract

This study scrutinizes to what degree an extension of the Capture-Recapture (CR) approach, extrapolated from ecology, gives us further insight into second language (L2) productive lexical desity. Previous CR studies that have adopted this method to estimate L2productive vocabulary size have typically taken only two written samples, produced within a few days of each other, from the participants. Results have invariably shown that this approach does not measure absolute vocabulary size. Another issue that compromises the CR scores is the kind of task used to elicit samples of productive vocabulary. In ecology there is an assumption that all animals of the target population have an equal chance of being captured. In a lexical context this assumption would translate as that all vocabulary items have an equal chance of being produced. This assumption presents a serious problem to overcome because different lexical items have different probabilities of being elicited.

* Mahidol University International College (MUIC)

This paper addresses both issues. On the one hand, three samples from 26 Thai students of Spanish as L2 have been taken within a period of two weeks (referred to as “an extension of the CR method”). On the other hand, a context-free task based on the Roman alphabet was used to elicit productive vocabulary. The results of this approach were compared to those of the CR method (using only the first two samples). The overall estimates are higher when using three samples. However, once again, the scores do not give absolute vocabulary size. A discussion of the method’s assumptions is provided together with an interpretation of the results.

# ECOLOGICAL APPROACH TO VOCABULARY DENSITY ESTIMATION

Dr. Juan Carlos Olmos Alcoy*

## Background

This paper intends to elaborate on the Capture-Recapture (CR) approach which has been recently used to estimate productive vocabulary size (Meara and Olmos Alcoy, 2010; Williams et al, 2014). The CR approach is a model originally developed in ecology in order to reliably estimate animal abundance (N) in particular habitats of interest. This is how the CR basic methodology works: on Day 1 a sample of the target population is captured, tagged or marked in some way for later identification, the number of individual is noted (T1) and finally released into the habitat where they came from. After a few days, when the captured individuals have had time to naturally distribute themselves in their habitat, a second sample is taken. This time we obtain two figures: the total number of individuals (T2) and the number of individuals captured on both sampling occasions (R). From this data a formula known as the Peterson Estimate (Peterson,

* Mahidol University International College (MUIC)

1896, Seber, 1973) is derived: N/T1 = T2/R. We obtain the N value by altering the formula as follows N= T1 x T2 / R. Of course, this formula assumes that the ratio of R with regard to T2 is the same as that of T1 with regard to N. In other words, if the value of R is, say, 25% of T2, then the value of T1 should be 25% of N too. However, in ecology, this holds true only if three assumptions are met; 1) the size of the target population between T1 and T2 remains constant, 2) the sample is random, and 3) all individuals have the same chance of being captured. If we extrapolate these three assumptions into a lexical context, this is what we obtain: 1) the amount of vocabulary students know between the completion of Tasks 1 and 2 remains constant, 2) lexical items are randomly produced, and 3) all lexical items are equally eligible for elicitation. The challenge is, then, devising a task that elicits productive vocabulary and meets the three assumptions afore-mentioned.

## Measuring productive vocabulary

Past attempts at measuring productive lexical density as L2 have conventionally used one of the following tasks in order to elicit vocabulary samples from the learners: a) narrative texts, b) word association lists, or c) tasks that elicit contrived responses from the participants.

We will now discuss to what extent the three assumptions mentioned in the previous section hold true in each one of the productive vocabulary elicitation tasks commonly deployed to extract L2 lexical productions.

**a) Assumption 1: the amount of vocabulary remains constant between T1 and T2.** This assumption is very likely to hold true because the time span between the completion of both tasks is practically insignificant. Even in the case that the participants are exposed to a lot of new lexical items just prior to T2, it is unlikely that this vocabulary is ready for productive use. When L2 learners come across new words, they normally acquire some receptive knowledge first, that is, the students might be able to recognize these new words if they read or listen to them again. However, on average L2 learners have to encounter receptively these words up to 7 times (Nation, 2001) before they can be used confidently in speaking or writing, that is, in a productive context. Hence, it is very unlikely that between the completion of T1 and T2 –typically 1 or 2 weeks apart- any significant acquisition of productive lexical knowledge has taken place.

**b) Assumption 2: lexical items are randomly produced.** Narrative tasks invariably require either a written or an oral production from the participants. In order to do so, they have to be given some topic – often an essay question, a descriptive task or a picture is used as stimulus to trigger a story. Consequently, narrative tasks create a lexical bias imposed by the topic itself. Writing about any specific topic inevitably facilitates the production of certain words and, at the same time, inhibits the production of many others. Even not very advanced participants will be productively familiar with lexical fields such us food and drinks, everyday habits, hobbies, sports, jobs, etc. but will be unable to use many of them because they are irrelevant. This means that the choice of topic “potentially ha(s) a very meaningful impact on the lexical richness of the text” (Fitzpatrick, 2000, p. 25).

Word association tasks are, to some extent, better suited to elicit lexical items randomly than narrative tasks. To begin with, as Meara and Fitzpatrick (1999) quite rightly state “a large number of vocabulary areas are opened up very economically” (p. 26); secondly, due to the irrelevance of grammar, function words like articles or prepositions – which tend to be very frequent and have no lexical meaning- are not necessary. This eliminates the strong element of repetition inherent in any narrative text. Test takers, hence, need to focus exclusively on lexical production without having to worry about grammatical accuracy. As a result, most items –if not all- elicited by word association lists are likely to be different. This is clever because it optimizes the participants’ chances to produce a rich and varied vocabulary within the set time limit.

There are, however, a few limitations with word association tasks. To begin with, their inherently associative nature prevents the participants from producing lexical items randomly. That is, when L2 learners complete the same task for the second time (T2), it is likely they use many of the items they used previously in T1. The assumption, then, that all words are randomly produced is seriously compromised. Also, there is an explicit expectation that (nearly all) participants are able to recognize all –or most of- the stimuli items. Hence -because of this recognition- they will be able to make associations. In fact, recognizing any of the stimuli items is not a necessary requirement to make associations. Let us look at the following theoretical -but certainly not implausible- examples of word associations:

Dig................ pig, piss, pill
Habit.............. rabbit, bandit, market
Rice............... mice, mile, mine
Beat............... meet, meat, eat

All these associations are morphologically based.  It is not clear whether the stimuli words are recognized or not. Furthermore, we cannot either be sure the participants know the meaning of the associations they write. Participants might have encountered some of them previously in class or in a book, but only because of the similar spelling, they are able to regurgitate these associations.

Lastly, highly contrived responses (Laufer, B. and Nation, P. 1999) force participants into eliciting specific lexical items. To avoid acceptable but unwanted answers, sometimes the first few letters of each target words are provided. The following is an example:

**The book covers a series of isolated epis____ from history.**

This type of productive test is heavily contextualized. This contextualization coupled with the fact that the first few letters of each item are provided, certainly guarantees that no random answers are produced.  It is impossible, then, to meet this second assumption with this type of task.

c) Assumption 3: all lexical items are equally eligible for elicitation. This assumption is also severely violated because words are divided into frequency bands. Inevitably, some words are much more commonly used than others. In narrative tasks we will find the same function words appear very frequently because they are necessary to make the text grammatically correct. Word association lists do away with this grammatical dimension. They mostly elicit nouns and adjectives –and, to a lesser extent, adverbs (Meara, P. and Fitzpatrick, T. 1999)- which are stratified by frequency bands. This means that many words will appear much more often than others, betraying the assumption that all lexical items are equally eligible for elicitation. Lastly, in the tasks devised to elicit contrived responses, it goes without saying that asking the participants to write pre-selected target items makes it impossible for this third assumption to be met.

## Summary

We have just seen why, so far, all tasks designed to elicit productive vocabulary from L2 learners fail to meet one or more of the three assumptions afore-mentioned. It is necessary for all these assumptions to hold true in order to achieve reliable conclusions. In the next section we will look into a different type of task designed to improve on some of the limitations that have pestered the measurement of productive vocabulary density. The test-takers will also be required to take this task 3 times (T1, T2 and T3) over a very short period of time.

## Methodology

### Participants

A total of 26 students from two different proficiency levels participated in this study. 14 students have studied Spanish as L2 for two academic years and were considered to have an intermediate level of proficiency. The other 12 students have studied Spanish for a period of over 3 years and were considered to have a (near) advanced level of proficiency.

### Data collection

The alphabet is presented in a column. This 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 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.*

Normally there is one letter of the alphabet per row. In some cases, two or more letters are grouped: comparatively speaking, in

Spanish there are not as many words beginning with J, K, N, W, X, Y or Z. For this reason in one row they can write words beginning with J or K, in another row they can write words beginning with N or Ñ, in another row we have V or W, and in the last row, they are asked to write words beginning with X, Y or Z. This is hoped to increase the students’ chances of production. Also, it can potentially diminish levels of frustration at not being able to produce, say, five words beginning with Ñ.

Since this task is completely decontextualized, it is also expected to elicit vocabulary items from a wide variety of lexical fields. The task is completed 3 times (T1, T2 and T3) over a period of two weeks. It is specified that:

- they should write the first word that comes to mind when they read each letter, and
- there are no right or wrong answers.

When the task is repeated, students are reminded that this is not a memory exercise and, therefore, should not attempt to recall past responses.

## Counting words

The following criteria are applied when counting words:

- 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 when scoring takes place. In linguistics this is called a lemma.

- Different instances of the same regular verb (i.e: hablo, hablaba, hablé) are also counted as one entry when scoring is done.
- Different instances of the same irregular verbs (i.e: voy, fuimos, vaya) are, however, counted as different entries. It was decided to treat irregular forms differently from regular forms because there is evidence to suggest that the brain deals with both types in different ways[1] .
- Proper nouns, numbers, abbreviations and function words are ignored when counting is done.
- Minor spelling mistakes are corrected and taken into account when counting is performed.
- Items which are not recognized as an existing Spanish word (i.e: because they are severely misspelled) are ignored.

[1] An important PET (*Positron Emission Tomography*) study scrutinized what areas of the brain were activated by regular verb past tenses versus irregular ones in English (Jaeger et al., 1996). It was found that "the irregular past task elicited significantly larger areas of brain activity at higher degrees of significance than did the regular past task… These findings strongly support the … hypothesis that regular and irregular past tenses in English are computed by different mechanisms" (p. 488). Similar results have been found in other languages like German (Beretta et al., 2003) and Spanish (Rodríguez-Fornells et al., 2002).

## Results

After the students complete the task for the third time, words are counted and the extension of the CR method is applied. The SPSS statistical program (Statistical Package for Social Sciences v. 17.0) was used to analyze the data obtained from the students. An alpha level of .05 was chosen as the significant level. Table 1 shows the overall number of words and the estimates of each group.

| | | Overall number of words used per group | Estimate (N estimate) |
|---|---|---|---|
| Int. | Mean | 192.14 | 263 |
| | *sd* | *23.15* | *61.86* |
| Adv. | Mean | 221 | 314.42 |
| | *sd* | *12.5* | *32.69* |

**Table 1**: Mean results for the Intermediate and Advance groups

A univariate analysis of variance is carried out between both groups. It shows a significant difference between both groups (F(1, 24)= 38.31, p < 0.01). This confirms that the Advance group knows significantly more productive words than the Intermediate group. An independent samples t-test also confirms that there is a significant difference between both groups, t(20.30) = -2.70, p =.014.

Table 2 shows the mean number of new words that the two groups generated for each of the three collection times and the number of repeated words in Tasks 2 and 3.

| | | Number of new words in all tasks | | | Number of repetitions in Tasks 2 and 3 | |
|---|---|---|---|---|---|---|
| | | Task 1 | Task 2 | Task 3 | Task 2 | Task 3 |
| Int. group | Mean | 92.86 | 58.71 | 42.71 | 37.71 | 55.14 |
| | *sd* | *9.08* | *10.96* | *8.32* | *8.34* | *8.08* |
| Adv. group | Mean | 101.17 | 69.25 | 50.58 | 37.17 | 55.26 |
| | *sd* | *7.88* | *6.00* | *7.29* | *7.22* | *6.84* |

Table 2: Mean results of the Intermediate and Advanced groups

These data is not difficult to interpret. Regarding the number of new words produced by both groups in each task, the difference is what we would have expected: the Advanced group, who has more lexical knowledge at their disposal than the Intermediate group, produces more new words in Tasks 2 and 3. Independent samples t-tests confirm that the difference is significant in Task 1 t(23.99) = - 2.49, p = .020, in Task 2 t(20.69) = -3.09, p = .006 and also in Task 3 t(23.98) = -2.56, p = .017.

Rather surprisingly though, both groups repeat nearly exactly the same amount of words in Tasks 2 and 3. An independent samples t–test shows there is no significant difference between these groups.

## Discussion

The results reported above raise a number of important issues that need to be addressed. The issues are a) the extent to which the extension of the CR method is a better estimator than the Peterson estimate, b) implications of the different patterns of responses, and c) the suitability of the alphabetical list as an eliciting task.

### a) the extent to which the extension of the CR method is a better estimator than the Peterson estimate.

The results reported above imply that this approach is once again partially successful. Furthermore, most N estimates are (remarkably) higher than those of the previous experiment (Meara and Olmos Alcoy, 2010). This suggests we may be another step closer to achieving more realistic figures. The mean N estimate for the Intermediate group is about 263 words, and for the Advanced group is about 315 words. The highest score, however, was reached by a student from the Intermediate group: 390 words. This is a considerable improvement when compared to most of the previous estimations. If we now apply the Peterson estimate (using Tasks 1 and 2) to the data, we can see how these two estimates compare.

Table 3 and Table 4 below provide a summary of the results:

| | Peterson estimate | Extension of CR |
|---|---|---|
| Mean | 253.5 | 263 |
| *sd* | *87.88* | *61.86* |

**Table 3**: Summary of results for the Intermediate group

| | Peterson estimate | Extension of CR |
|---|---|---|
| Mean | 297.1 | 314.42 |
| *sd* | *47.82* | *32.69* |

**Table 4**: Summary of results for the Advanced group

We will now assess to what extent the three assumptions mentioned before are met.

**Assumption 1**: the number of words students know between the completion of Tasks 1, 2 and 3 remains constant. This assumption is very likely to hold for both groups because all three tasks were completed in a very short period of time (2 weeks). It is assumed that in this period of time no significant level of lexical learning or attrition took place.

**Assumption 2**: words are randomly produced. The amount of repetitions produced by each group gives us some indication of how random responses are. It is theorized that the more new words occur, the more likely responses are randomly elicited. Clearly, the advanced group elicits more new items than the Intermediate group. This can be appreciated in Table 2. It confirms that the Advanced students have significantly greater lexical knowledge at their disposal; however, we should not stop here. The task was designed to maximize 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 (i.e: brebaje, umbral, relámpago, ubicuidad). This feedback suggests that to some extent random elicitation was achieved. Interestingly, this type of feedback also contrasts to a degree with Laufer’s definition of use (2002); she specifies that if students use a word correctly in context in their written or spoken productions, then we can say they have use of the word. Ironically, in this study the lack of context is what has propitiated the use of particular words. It is not immediately clear why this is the case. Strictly speaking, we cannot say that these words show evidence of use 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.

Assumption 3: all words are equally eligible for elicitation. We know, however, this is not the case: most words have a different likelihood of being produced because they are affected by their frequency.

b) Different pattern of responses.

In Table 2 we can see the means given by both groups in a) the number of new words in all tasks and b) the repetitions in Tasks 2 and 3. They are going to be discussed in turn now.

In Table 2 we can, on the one hand, appreciate the number of new words used as students repeat the task. It is not surprising that the Advanced group always produces more new words than their Intermediate counterparts. In Tasks 2 and 3 the number of new words decreases noticeably. This decline is slightly sharper in the Intermediate group. This suggests that the Advanced students have more words at their disposal. Intriguingly, both groups display virtually the same quantity of repetitions in Tasks 2 and 3. It was hoped that the more proficient students would produce fewer repeats than their less proficient counterparts but this was not the case. It is not clear why this is the case and further experimentation should be carried out to see if the same pattern emerges with other groups.

The stimuli items –the alphabet- influences both these patterns though. Some letters, like C, M or S, seem to encourage diversity of responses: there are many words in Spanish beginning with these particular letters. Conversely, letters like J, K or LL often elicit the same lexical items (i.e: jamon, jardín, kilo, llamar, lluvia). Repetitions elicited by these letters can be appreciated not only in the responses given by the same student in all tasks; different students often repeat the same items as well. This suggests that randomness is compromised to a degree: these letters threaten spontaneous lexical production.

Instead, they tend to compel students to write certain items. In future experiments of this kind, it may be worth considering the exclusion of these letters in order to optimize random elicitation of vocabulary.

Due to the constraints of the task, the maximum number of responses on any particular occasion T1, T2 or T3 is 115, this is the maximum number of items that could possibly be elicited. This implies that some students would have been able to write (many) more responses had the task required bigger productions. In practice, the Intermediate group produced about 93 words on average and the Advanced group produced 101 words. In both groups very few responses were either ignored when counting was done and/or some responses were counted as one. Few items were severely misspelled (i.e: rampo, lorrar, lampos; maybe rompo, llorar and campos are meant but we cannot be sure), not Spanish (i.e: french, nonsensico) or (seemingly) made up (i.e: zizage??). On occasion several instances of the same verb appear (i.e: leer, leyendo; fui, fue, fuiste) but, again, this was uncommon. Students produced mostly nouns and verbs belonging to a large spectrum of lexical fields. Some of these include: parts of the body, food and drink, directions, members of the family, furniture and kitchen utensils, holidays, animals, colors, sports and games, parts of the house, feelings, jobs, studies, geography, means of transport, weather, clothes, buildings, etc. Adjectives also appear often but much less frequently than nouns and verbs. Very few adverbs and function words are elicited.

Since the students have complete freedom to write any items they like, it is not surprising that both groups reach very similar levels of production. Table 5 shows the mean number of words each group produces in each task.

| | | Task 1 | Task 2 | Task 3 |
|---|---|---|---|---|
| Int group | Mean | 92.86 | 96.42 | 98 |
| | *sd* | *9.08* | *7.92* | *9.27* |
| Adv group | Mean | 101.17 | 105.75 | 105.83 |
| | *sd* | *7.88* | *6.56* | *5.62* |

Table 5: Mean number of words produced in each task

We should be very cautious with any interpretation of the data in Table 5 because we only have 3 mean values per group; still, we can tentatively discern two slightly different patterns of responses. The level of participation of the advanced group increases from Task 1 to Task 2 and then remains virtually unchanged in Task 3. The Intermediate students show a small but steady increase every time they complete the task. This pattern may be due to a combination of two factors:

- They were already familiar with the task and found it slightly easier to elicit a few more items every time they completed the task and/or

- Fewer items were deleted or grouped together as one lemma (i.e: amigo, amiga, amigos, amigas = 1 word) before counting took place.

The fact that both groups reached a high level of production in all tasks suggests that they could have written (many) more responses, had they been requested to fill in more columns. This has a danger though: the longer the productions, the more likely repetitions will increase (Arnaud, 1984; Laufer and Nation, 1995; Malvern and Richards, 1997). This is something we ideally want to avoid because repetitions –the denominator of the formula in both the Peterson estimate and the extension of the CR method- have a (very) detrimental effect on the N estimate.

c) Suitability of the alphabetical list as an eliciting task.

It was theorized that the use of the alphabet as a prompt to stimulate lexical production would increase randomness and decrease repetition. The results suggest that we have achieved both goals to a degree.

Let us remember that ensuring randomness is vital in all ecological models in order to avoid biased results (Krebs, 1999). Using the alphabet to stimulate lexical production seems to promote randomness better than pictures (Meara & Olmos Alcoy, 2010). These are two factors that suggest this:

- Feedback given by the students indicates that many of their answers tend to be spontaneous, and
- The responses produced in each task come from a wide range of lexical fields.

The lowest number of repetitions is 21. If we take this value as representative and apply it to both groups, we obtain the following figures:

| | | Extension of CR method (N estimate) |
|---|---|---|
| Int group | Mean | 1137.21 |
| | *sd* | *201.77* |
| Advanced group | Mean | 1368.5 |
| | *sd* | *145.78* |

Table 6: Mean results of the Intermediate and Advanced groups (with 21 repetitions)

These results are much closer to the estimates we ideally want to reach. Not surprisingly, an independent samples t-test also confirms both groups are significantly different t(23.39) = -3.38, p = .003. Lastly, although the N estimates are still low in terms of overall lexical density, we can appreciate a marked improvement when compared to previous estimates.

## Conclusion

In this paper we have seen how an extension of the CR method can be used to estimate productive vocabulary size. We have also used the alphabet as stimulus to optimize random lexical elicitation as well as minimize repetitions. Results are encouraging to a degree because the N estimates are higher than those of previous experiments. This suggests that we are on the right track because we are getting closer to finding an approach that can give us realistic estimates. Future experiments should explore other approaches and methodologies in order to further enhance our N estimates.

## References

Arnaud, P. (1984). The Lexical Richness of L2 **Written Productions and the Validity of Vocabulary Tests**. University of Essex: Dept of Language and Linguistics Occasional Papers No. 29, pp. 14-28.

Beretta, A., Campbell, C., Carr, T. H., Huang, J., Schmitt, L. M., Christianson, K., & Cao, Y. (2003). An ER-fMRI investigation of morphological inflection in German reveals that the brain makes a distinction between regular and irregular forms. **Brain and Language**, 85 (1), pp 67-92.

Fitzpatrick, T. (2000). Using Word Association Techniques to Measure Productive Vocabulary in a Second Language. **Language Testing Update**, 27, 64-70.

Jaeger, J. J., Lockwood, A. H., Kemmerer, D. L., van Valin, R. D. Jr., Murphy, B. W., & Khalak, H. G. (1996). A Positron Emission Tomographic Study of Regular and Irregular Verb Morphology in English. **Language**, 72 (3), pp 451-497.

Krebbs, J. C. (1999). **Ecological Methodology.** Addison-Welsey Educational Publishers, Inc.

Laufer, B., Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written productions. Applied **Linguistics** 16: 307-322.

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

Laufer, B. (2002a). Computer Adaptive Test of Size and Strength, Paper presented at Second Language Vocabulary Acquisition Colloquium, Leiden, March 2002.

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: Multilingual Matters.

Meara, P., & Fitzpatrick, T. (1999). Lex30: an improved method of assessing productive vocabulary in an L2. *System 28*. pp. 19-30.

Meara, P., and Olmos Alcoy, J. C., (2010). Words as species: An alternative approach to estimating productive vocabulary size. **Reading in a Foreign Language**, 22. pp. 222-236.

Nation, I. S. P. (2001). Learning Vocabulary in Another Language, Cambridge University Press.

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

Rodríguez-Fornells, A., Munte, T. F., & Clahsen, H. (2002). Morphological Priming in Spanish verb forms: an ERP repetition priming study. J**ournal of Cognitive Neuroscience**, 14 (3), pp 443-454.

Seber, G. A. F. (1973). **The estimation of animal abundance and related parameters.** Bristol: J. W. Arrowsmith Ltd.

Williams, J., Segalowitz, N., & Leclair, T (2014). Estimating second language productive vocabulary size: A capture-recapture approach. **The Mental Lexicon**, 9 (1), 23-47.