Bayesian Vocabulary Tests

Paul M. Meara
Lognostics, Cardiff, UK
p.m.meara@gmail.com

Imma Miralpeix
Department of English Studies, Universitat de Barcelona, Spain
imiralpeix@ub.edu

Abstract

This paper proposes a new way of looking at productive vocabulary in L1 and L2 speakers. An experiment was conducted where 160 participants provided six words for five different picture prompts they were presented with. Data from this minimal vocabulary test was analysed using Bayesian statistics in order to decide whether a set of responses were generated by an L1 speaker or by an L2 advanced learner. Results obtained provide some interesting insights into the viability of minimal vocabulary tests (small sets of words can carry large amounts of information on vocabulary use), as well as some indications of how Bayesian methods could help us explore productive vocabularies of L2 speakers at different proficiency levels.

Keywords: Bayesian statistics, L2 learning, minimal vocabulary tests, productive vocabulary, vocabulary testing.

Resumen

Este artículo propone una nueva forma de considerar el vocabulario productivo en hablantes nativos y aprendices de segundas lenguas. Se realizó un experimento en el que 160 participantes proporcionaron seis palabras para cinco imágenes diferentes que se les presentaron. Los datos de esta prueba mínima de vocabulario se analizaron utilizando estadística bayesiana para decidir si un conjunto de respuestas fue generado por un hablante nativo o por un aprendiz de nivel avanzado. Los resultados obtenidos ofrecen ideas interesantes sobre la viabilidad de las pruebas mínimas de vocabulario (un pequeño número de palabras puede proporcionar gran cantidad de información sobre el uso del vocabulario), así como algunas indicaciones de cómo los métodos

bayesianos podrían ayudarnos a explorar los vocabularios productivos de hablantes de idiomas a distintos niveles de competencia.

Palabras clave: Estadística bayesiana, aprendizaje de segundas lenguas, pruebas mínimas de vocabulario, vocabulario productivo, test de vocabulario.

1. Introduction

In L2 vocabulary research, a distinction between receptive and productive vocabulary knowledge has often been made: we usually assume that receptive vocabulary involves being able to recognize and understand a word when it is encountered in listening or reading, while productive vocabulary means being able to use it in speech or writing. There is also a general agreement by the research community that receptive vocabularies tend to be bigger than productive vocabularies, as reception precedes production (e.g. see Melka, 1997; Webb, 2008). Receptive vocabulary size has been object of study for a very long time and several tests have been proposed to measure this dimension, namely multiple choice tests, for example the *Vocabulary Levels Test* (VLT: Nation, 1990; Schmitt, Schmitt, & Clapham, 2001; Webb, Sasao, & Balance, 2017) and the *Vocabulary Size Test* (VST: Nation & Beglar 2007; Coxhead, Nation, & Sim, 2014) or yes/no tests (Meara & Buxton, 1987) such as *V_YesNo* (Meara & Miralpeix, 2015b). However, productive vocabulary size has been largely unexplored, which is partly due to the need for new assessment methods. Different approaches have been taken in an attempt to measure productive vocabulary size. So far, (1) we are able to describe the sort of vocabulary learners produce in a speaking or writing task; (2) we can measure 'controlled productive vocabulary size' when learners are asked to provide a specific word given its first letters; (3) we can derive scores from word associations tests or lexical availability tasks that may give an idea of how big vocabularies are and (4) we can use mathematical methods typical of other fields, such as biology, to estimate the amount of words students of a language may know productively. Nevertheless, these approaches have also raised several concerns, as we will see in sections 1.1-1.5 below. Therefore, in section 2 we propose a new method to explore productive vocabulary sizes following Bayes' theorem. Our aim in this paper is to evaluate the Bayesian approach by analysing its potential for distinguishing between L1 and L2 speakers on the basis of very few words, which were produced to describe a set of picture prompts (sections 3-5).

1.1. The Lexical Frequency Profile

One of the first attempts to characterise learners' productive vocabulary is Laufer and Nation's Lexical Frequency Profile -LFP- (Laufer & Nation, 1995). The operation

of the LFP is essentially very simple: LFP takes a raw text as input and returns as output a profile of the text in terms of the frequency distribution of its words. Laufer and Nation suggest that a profile based on four frequency categories is useful - the four categories being based on Nation's earlier work on word lists for L2 learners (Nation, 1984). Category 1 consists of the 1000 most frequent words in English as defined by Nation's lists; category 2 consists in the second 1000 most frequent words; category 3 consists of words in the University Word List (Xue & Nation, 1984); category 4 includes any word not found in the previous three lists. It should be acknowledged, though, that Laufer and Nation's profiles are not particularly easy to work with: they describe a learner's output as a four point profile, rather than as a single measure, and it is difficult to summarise the data that they encapsulate in an economical and transparent way. Furthermore, rather than predicting testees' lexical proficiency, the LFP describes the kind of words testees use in any piece of oral or written data.

1.2. Controlled productive vocabulary

Laufer (1998) makes a distinction between controlled productive vocabulary and free productive vocabulary. She defined a test of controlled productive vocabulary as one that "entails producing words prompted by a task" (e.g. when the first two letters of a word in the context of a sentence are provided to the student), whereas free productive vocabulary "has to do with using words at one's free will, without any specific prompts for particular words" (1998: 257). Laufer and Nation (1999) developed a test to assess the former, by using the same words as in the VLT receptive vocabulary test. In this case, the test-takers responses are constrained by providing the first few letters of the expected response, as in the example below:

The house was su_____ by a big garden. *(surrounded)*

However, assessing free productive vocabulary is much more challenging, as research on the topic has clearly evidenced.

1.3. Vocabulary size from word association and lexical availability tasks

Although word association tests (Meara & Fitzpatrick, 2000) or lexical availability tasks (Roghani & Milton, 2017) were not first devised for this purpose, data from these tasks typically consist of L2 words that could be profiled using standard vocabulary assessment tools such as Range (Heatley, Nation & Coxhead, 2002). For example, in lexical availability tasks learners are asked to name as many words as they can from a prescribed category, such as *food*, *parts of the body*, *animals* or *transport* (Jiménez Catalán, 2014). The learner profiles obtained from learners' answers could provide a picture of

the scope of a testee's productive vocabulary, as shown above. However, more research is needed on the extent to which these profiles might be good indicators of productive vocabulary knowledge (Fitzpatrick & Clenton, 2017).

1.4. The capture-recapture method: V_Capture

V_Capture is a computer program based on the idea developed by biologists interested in counting the number of species in a particular area by capturing and then recapturing animals in traps on a number of different occasions. This process is similar to comparing the words produced in a particular task over several performances. Mathematics uses the proportion of animals captured (in the case of vocabulary that would equal words used) on both occasions to estimate the number of animals or species being studied (Meara & Olmos Alcoy, 2010). In spite of the fact that the program computes Petersen Estimates (and thus provides an estimate of vocabulary size), there are several problems with these estimates for continuous texts, and more plausible results can be obtained from wordlists rather than from texts (Meara & Miralpeix, 2017). Interesting work on the capture-recapture method can also be found in Williams, Segalowitz and Leclair (2014).

1.5. A productive vocabulary size estimator: V_Size

As it is problematic to assess total productive vocabulary size, it may be more suitable to compare relative vocabulary sizes, such as the vocabulary someone uses for a particular task compared to what others (e.g., NS or learners at different proficiency levels) use when performing the same tasks. A range of different tasks, such as cartoon storytelling or picture description, may be needed for this purpose, and tools using different estimation methods can help researchers obtain reliable estimates. Estimates by *V_Size* (Meara & Miralpeix, 2015a) are based on the Power Law, a ranked distribution found not just in language but also in other physical and biological phenomena like earthquake size or social network connectivity. The program assumes that certain words in language (in high frequency bands) are more frequent than others (in low frequency bands) and that there is a direct relationship between the number of times a word occurs in a corpus and its rank in a frequency list generated by the corpus. Thus, it allows researchers to go beyond the mere shape of the frequency profile generated by a text and enquires into what the profile tells us about the size of the productive vocabulary of the person who produced the text. As noted by Castañeda-Jiménez and Jarvis (2014: 501), *V_Size* "is the only freely available computer program we [the authors] know of that outputs estimates of learners' productive vocabulary based on the texts they produce".

While *V_Size* may give us good indications in the future of the vocabulary known productively for certain tasks, it would also be very useful for us to determine learners' proficiency level from a sample of words they know productively. A first step in this direction would be trying to distinguish between native speakers (NS) and advanced learners on the basis of very few words, which is what we will try to do in this paper.

2. Bayes theorem and its applicability to vocabulary assessment

As noted by Miralpeix (2020), among others, assessing productive vocabulary always involves eliciting a set of words from learners and inferring from this sample the size of the learners' repertoire, i.e. how many words they would be able to retrieve from memory without seeing them written or hearing the spoken forms. All estimations are based on probabilities, as it is impossible to elicit from learners all the words they know productively in an L2 (unless they are at the very first stages of learning a language and know very few words).

Up to now, the mathematical procedures that we have used for productive vocabulary measurement have relied on analysing learners' data using proportions (e.g. in the capture-recapture method, see section 1.4) or comparisons of rank distributions with curve-fitting (e.g. in *V_Size*, see section 1.5). It has also been observed that the more a learner speaks (or writes), the more capable we are of making inferences about his/her lexical knowledge, as every word introduces new information that can help us make guesses about his/her level. It would be really useful if we could formalise guesses of this type and one way of doing this is to apply Bayesian statistics to the data.

The immediate background to the work we present in this paper is an earlier study (Meara & Miralpeix, 2017) in which we asked L1 Spanish and Catalan learners of English to generate a set of ten adjectives in response to a cartoon stimulus like the one shown in Figure 1. This apparently simple task turned out to be quite difficult, even for advanced learners.

Figure 1: The cartoon figure used in Meara & Miralpeix (2017).

The main focus of our research at the time was how similar learners' response sets were. We will not discuss this work here other than to say that the cartoons generated a very large number of disparate responses, and that a response set generated by advanced L2 speakers would typically share just over two words with another response set from the same group of participants. Some examples are provided in Table 1. These responses were made by first language Catalan or Spanish learners of English.

Table 1: Some example response sets made to Figure 1.

NNS001,slim,intelligent,serious,tall,big_headed,angry,lonely,bad_tempered,strict,helpful
NNS002,smart,formal,strict,serious,slim,bad_tempered,intelligent,thoughtful,impatient,rude
NNS003,big_headed,tall,ugly,slim,bald,serious,shy,open_minded,young,impartial
NNS004,ugly,short,big_headed,evil,strange,bad_tempered,scary,angry,lonely,moody
NNS005,grumpy,surprised,big_headed,lunatic,creepy,dirty,stinky,weird,adult,ugly

While working with these data, we noticed that we were often able to decide whether a set of responses was generated by an L2 learner or an L1 speaker. Obviously, the L2 speakers sometimes produced responses that were easily identifiable as learner

errors, but leaving responses of this kind aside, we noticed that some words were more likely to be used by learners than by native speakers, and vice versa. Some words were almost exclusively used by one group rather than the other, while other words were somewhat more likely to be used by one group. For the stimulus picture in Figure 1, the two groups generated a total of 650 different words, which can be divided into three categories: words used by both groups of participants (shared words: 30%), words used predominantly by the L1 participants (L1 words: 46%) and words used predominantly by the L2 participants (L2 words: 24%). With practice, one becomes fairly good at guessing whether a data set was generated by an L1 speaker or and L2 learner.

One way to formalise this intuition is by means of Bayes' Rule (McGrayne 2011). Bayes' Rule is a mathematical procedure which allows us to change our estimate of something being true, in the light of additional evidence. This approach has not been popular in language acquisition studies, although recently Norouzian et al. (2018) introduced the application of Bayesian methods to various research designs, and Pearl and Goldwater (2016) and Zinszer et al. (2018) used Bayesian inference models to analyse L1 acquisition. Bayes theorem has mostly been used in situations where data is partial and difficult to interpret - naval searches, medical diagnoses, face recognition and spam filtering, to name but a few. This last example is particularly interesting for us, as the best spam filters rely on a comparison of the words typically used in spam emails, and the words used in bona fide emails - a comparison that is not a million miles away from the problem we are faced with when we try to assess the vocabularies of L2 speakers.

The approach is usually described as follows:

$$P(A|B) = [P(B|A)\ P(A)] / P(B)$$

Which tells us:	how often event A happens *given that B happens*, written **P(A|B)**,
When we know:	how often event B happens *given that A happens*, written **P(B|A)**
	and how likely A is on its own, written **P(A)**
	and how likely B is on its own, written **P(B)**

Although the approach has not been used in vocabulary assessment, Bayes' rule could help us predict proficiency levels from peoples' productive vocabularies. For instance, it can give us information on the chance a set of words being really from an L2 learner or a NS taking into account the words that they produce in a test, as there are words with higher chances of appearing in L2 learners' sets than in NS sets. In this case:

Our Event A: The set is produced by an L2 learner
Our Event B: The presence of certain words

Then:
P (L2 | words) = [P(words | L2) P(L2)] / P(words)

By making this filtering, based on updating probabilities, we could know, for example, if a set has an 85% chance of being produced by an L2 learner (then, it probably is) or if it has a 10% (then, it probably has been produced by a NS).

3. The study

3.1. Research question

In the light of the previous research on measuring productive vocabularies, it would be interesting to explore the potential of Bayesian statistics to correctly identify different proficiency levels from a set of words provided by participants at these levels. In this paper, the research question that we try to answer is the following:

How can a Bayesian approach help in distinguishing between NS and advanced English learners from a small set of words they provided for the same picture stimuli?

3.2. Method

3.2.1. Participants

Two groups of 100 test participants were assembled: a group of 18-21 year-old L1 English students at Swansea University (mean age 20.1), and a group of L1 Spanish/Catalan students of the same age range (mean 20.3) at an advanced level at the University of Barcelona. The final sample for the present study consists of 160 participants (80 per group). In the NS group there were 53 females and 27 males, and 59 females and 21 males in the L2 learners' group. These learners were in the third year of English Studies, a degree on English Linguistics and Literature taught in English since the first year. At the moment of data collection, they had a C1 level and a 30% of the sample had been abroad to English speaking countries for a month or less. Although no placement test was conducted for the purpose of this study, other cohorts at this level have been shown to have receptive vocabulary sizes between 6,500-7,000 words.

3.2.2. Instruments

A set of fifty cartoons was commissioned in the same style as the cartoon that appears in Figure 1. Ten of the cartoons depicted older men, ten depicted older women, ten depicted younger men, ten depicted younger women and ten depicted young children. Five cartoon pictures were selected: an older man, an older woman, a young man, a young woman and a child. The aim for selecting these five (shown in Figure 2) was that they looked as different as possible so that testees had a high chance to provide enough words for each, without repeating any.

Figure 2: The stimulus pictures used in the study.

3.2.3. Procedure

Participants were given the picture set and asked to provide six adjectives that might describe the person in each picture (no examples or training were provided). They were asked to write the words on the dotted lines after the prompts *'Neville is ...'*, *'Margaret is...'*, etc.

It should be noted that, in some ways, this approach is similar to that followed in the productive tests we mentioned earlier (e.g. lexical availability tasks), in the sense that we will be working with relative vocabulary sizes, i.e. the vocabulary someone uses for a particular task compared to what others use when performing the same task. Therefore, the approach will also be limited in the insight it provides about vocabulary proficiency 'in general'.

Not all participants managed to provide six answers for all the pictures (the L1 Ss in particular often produced a phrase rather than the single word that was requested). From the original 200 Ss, we managed to construct two sets of 80 Ss who generated six words for each of the five pictures. These groups were divided into two. For each L1 we

set up a group of 50 Ss whose data was used to establish a reference file for the group. The remaining 30 Ss were set aside to be used an evaluation group.

Next, for each reference group, we identified all the word types generated in their responses, and from this raw data we were able to identify word types used by both groups (shared words), words which were used only by the L1 group, words which were used only by the L2 group and singletons which occurred only once in the data set.

The question we then ask is whether these data can reliably predict the provenance of a new response set. In order to test this idea, we used the Bayesian approach described above to evaluate the 60 response sets that were left out of the analysis – 30 L1 speakers and 30 L2 speakers. For each response set, we estimated the probability of its being produced by an L1 speaker. Probabilities greater than 0.6 were taken to indicate that the data were produced by an L1 speaker, while probabilities below 0.4 were deemed to indicate that the data comes from an L2 speaker. Data sets where the final probability lies between 0.4 and 0.6 are deemed to be undetermined.

3.2.4. Data analysis

As Bayesian statistics is probably unfamiliar to most readers of this journal, we will explain the approach in some detail using as an example a data set of 10 words:

weird, unpredictable, old, angry, clever, worried, intense, interested, kind, thoughtful

Given raw data from suitable groups of participants, we can draw up a table which shows what we can expect of an L1 participant and an L2 participant in this task. In this particular case, we would expect about 75% of the responses generated by L1 participants to be shared words, about 20% of their responses to be L1 words and perhaps one of their responses to be a typical L2 response. For L2 speakers, we would expect 82% of their responses to be shared words, about 13% of their responses to be typical L2 words and perhaps one of their responses to be a typical L1 word. These response patterns look fairly similar (See Table 2), but taken together, the differences are large enough to allow us to ascribe an individual data set to the L1 group or the L2 group with a fair degree of confidence.

Table 2: The composition of the L1 and L2 data sets

	Shared Responses	L1 Responses	L2 Responses
L1 participants	75%	20%	5%
L2 participants	82%	5%	13%

Figure 3 shows how this works in practice. A given response set can either belong to an L1 speaker or an L2 speaker. Assuming that we do not know which answer is correct at the very beginning, we first assign both outcomes a probability of .5, as we start from a 50/50 hypothesis. Next, we look at word 1 in the response set (which is the word **weird** in example) and carry out the calculations shown in Table 3.

Table 3: Recomputing a probability in the light of new data

1. Find the appropriate column in Table 2. As an example we take **weird**, from the data set above. As it is an L2 word, so we work with the figures in column 3.

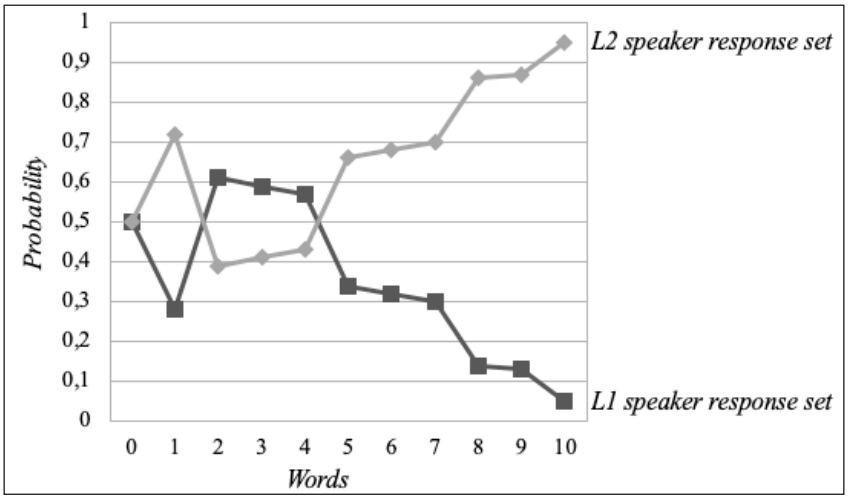
2. Multiply our current NS estimate by .05 .5*.05 = .025
3. Multiply our current NNS estimate by 0.13 .5*.13 = .065

4. Rescale the new probabilities so that they sum to 1. .025+.065 = .090

   new NS estimate: .025/.090 = .278
   new NNS estimate: .065/.090 = .722

Figure 3 shows, first of all, how the information provided by **weird** changes our assessment of whether this data set is generated by an L1 speaker or an L2 speaker: taking **weird** into account, it now seems slightly more likely that we are dealing with an L2 speaker.

Figure 3: Changes in confidence as more words are added to the data set



Next, we repeat the steps detailed in Table 3 using the new probabilities that resulted from step 3; that is, .278 (the new NS estimate) and .722 (the new NS estimate) instead of .5 (which was adopted for the first word in the data set). These steps are an implementation of *Bayes' Rule* (McGrayne, 2011; Stone, 2013). As our next word in the example set is **unpredictable**, which is a NS word, we use the probabilities in column 1 of Table 2. Applying the steps in Table 3, we get two new estimates: the L1 speaker estimate rises to .625 and the L2 speaker estimate falls to .375.

Finally, applying these steps to all ten of the words in the data set produces a convincing result: the probability that this response set is generated by an L2 participant is 0.95 (see Figure 3).

This is a pretty remarkable outcome. We started out with a mere sample of 10 words, generated to a simple cartoon, and we end up being 95% certain that the 10 words were generated by an L2 speaker. It is more than a little surprising that such a small data set can carry so much information, and allow us to make such confident assessments. It can also be observed in Figure 3 that by word 6 we start having a clear indication on whether the words were produced by an L1 or L2 speaker, that is why we opted for asking participants in the present study to provide six words for each stimulus. Therefore, these calculations were made using the six words in each of the sets that our participants produced.

3.3. Results

In this section we present the results obtained for each of the pictures used in the study. For each cartoon we provide some examples from the corpora we gathered, as well as the probabilities of response types according to the reference data (corpus), and a final assessment on the extent to which participants were classified as NS or learners using Bayesian statistics.

Table 4 shows the data for ***Shirley***, the cartoon of the young woman. Table 4a lists the words used to describe Shirley, divided into shared words, L1 responses, L2 responses and singleton responses.

Singleton responses make up a large proportion of the data (just over half the words fall into this category). Table 4b shows the probability of the different response types in the reference data set. Table 4c shows the way the test data are classified by the Shirley reference data set.

Table 4: Classified data for Shirley

Table 4a: Examples of words in the corpora for Shirley

Shared responses
SURPRISED NICE FREE ACTIVE SWEET CHEERFUL STYLISH THIN SINGLE POSITIVE WILD TALKATIVE LOVELY SEXY SMILING DANCER AFRO FRIENDLY HAIRY YOUNG CONFIDENT SKINNY FUNNY HAPPY SMILEY CURLY CHEEKY FASHIONABLE BLACK CRAZY PRETTY ENERGETIC SASSY TOOTHY DANCING EXCITED JOYFUL TALL FEMALE OUTGOING EXTROVERT COOL LIVELY

Typical L1 responses
ANNOYING DAME TEETH SINGER LOUD ENTHUSIASTIC JAZZY APPROACHABLE FUNKY HAIR SNAZZY BIG_HAIR BUBBLY SMUG TEETHY FUN

Typical L2 responses
BIG_MOUTHED HIGH POSH SHALLOW ATTRACTIVE SLENDER OPEN_MINDED UGLY SENSUAL GOOD_LOOKING SELF_CONFIDENT PLAYFUL FASHION CURLY_HAIRED SMART EXPRESSIVE ARTISTIC NERVOUS ELEGANT RICH SLIM BEAUTIFUL EXTROVERTED TRENDY WITTY

Table 4b: Probabilities of a specific response type in the Shirley reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.223	.213	.050	.514
L2 participants	.194	.050	.247	.509

Table 4c: Discrimination between the 60 response sets for the Shirley reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	23	6	1
L2 speaker response set	4	5	21

This cartoon is clearly very good at distinguishing between data sets generated by the two groups. Only one L1 participant is incorrectly classified, while 23 L1 speakers are correctly identified as such. For the L2 participants, the classifications are not quite so good: four L2 participants are incorrectly classified, but 21 are correctly identified as L2 speakers, with only five undecided. A chi squared analysis suggests that this distribution is very unlikely to have arisen by chance (χ^2=43.5; p<.001).

On the face of things, this looks like a very satisfactory result. Even when we simplify the test task by asking participants to produce only six words, the approach can correctly classify almost 75% of the response sets, and only one of the L1 speakers is misclassified. The L2 group contains a number of very high-level participants, and so we might expect the classifier to make some errors where an L2 speaker is judged to be performing like a L2 speaker on this task. The group of four L2 speakers who are misclassified seems like an allowable error.

Unfortunately, the results of the four remaining tasks are rather less compelling. Table 5 shows the data for ***Margaret*** – the older woman. The figures in Table 5b are quite close to the corresponding figures in Table 4b. The main difference is that both groups in the reference data set are about equally likely to generate one of the shared responses. This makes the classification rather more difficult, and Table 5c indicates that ***Margaret*** is indeed less good at discriminating between the groups than ***Shirley*** was. Here, 37 of the test cases were correctly identified as L1 or L2 speakers, but twelve L2 speakers were incorrectly classified as L1 speakers, and four L1 speakers were classed as an L2 speaker. Seven cases were undecided. A chi squared analysis suggests that this distribution is unlikely to have arisen by chance (χ^2=10.9; p<.01).

Table 5: Classified data for Margaret

Table 5a: The words participants use to describe Margaret

Shared responses
WAVING STUDIOUS FOREHEAD HARD_WORKING ENTHUSIASTIC ADULT SMILING CALM SENSIBLE QUIET HAPPY BLIND WELCOMING GLASSES GENEROUS INTELLIGENT NICE MOTHER SWEET OLD GENTLE CARING STRICT CHEERFUL FUNNY MIDDLE_AGED POSITIVE APPROACHABLE CLEVER WOMAN KIND SHY MOTHER_LIKE FRIENDLY SMART FEMALE CONSERVATIVE TEACHER

Typical L1 responses
BOOKWORM HAIR SMILEY WAVY GROOMED HELPFUL CASUAL DANCING KNOWLEDGEABLE PARTIALLY_SIGHTED KEEN LIBRARIAN CAT_LOVER PROFESSIONAL INNOCENT LOVING SIMPLE MOTHERLY

Typical L2 responses
SENSITIVE WISE, SHORT_SIGHTED PATIENT NERVOUS MATURE LOVELY RELAXED TALKATIVE EMPATHIC NAIVE EASY_GOING SHORT BEAUTIFUL OLD_FASHIONED CHARMING STRAIGHTFORWARD SYMPATHETIC POLITE FAMILIAR CURIOUS RELIABLE PRETTY MIDDLE_AGE OPEN_MINDED WELL_MANNERED RESPONSIBLE HONEST

Table 5b: Probabilities of a specific response type in the Margaret reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.233	.166	.050	.551
L2 participants	.216	.050	.283	.451

Table 5c: Discrimination between the 60 response sets for the Margaret reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	21	5	4
L2 speaker response set	12	2	16

Neville (Table 6) was fairly good at classifying the data sets, but identified a large proportion of undecided cases (15), and incorrectly classified five L1 speakers and seven

L2 speakers. Again, the overall distribution was not likely to have occurred by chance (χ^2=10.7; p<.005), but the relatively large number of undecided cases, particularly for the L1 speakers, is a problem.

Table 6: Classified data for Neville

Table 6a: Examples of words in the corpora for Neville

Shared responses
MAN MOUSTACHE TIRED SERIOUS MALE FUNNY ANGRY BORED MIDDLE_AGED STRICT SHORT RETIRED ARROGANT BOLD BIG_HEADED LONELY OLD_FASHIONED SHY CREEPY BAD ELDERLY WELL_DRESSED OLD WEALTHY NARROW_MINDED HAIRY RICH WISE CONCERNED GRUMPY CLEVER TRADITIONAL SLEEPY INTELLIGENT INTIMIDATING BALD TEACHER SAD NICE THOUGHTFUL BORING MARRIED

Typical L1 responses
PROPER OLDER QUIET BUSHY PROFESSIONAL STERN BIG_HEAD AWKWARD SUIT BALDING MISERABLE INQUISITIVE EYEBROWS GRANDAD RUSSIAN SMART SNOBBY

Typical L2 responses
UNFRIENDLY DISTANT RESPECTFUL RESPONSIBLE RUDE ELEGANT FAT UGLY EXHAUSTED CLOSE_MINDED THINKING WEIRD FORMAL BAD_TEMPERED

Table 6b: Probabilities of a specific response type in the Neville reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.250	.206	.050	.494
L2 participants	.183	.050	.140	.627

Table 6c: Discrimination between the 60 response sets for the Neville reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	16	9	5
L2 speaker response set	7	6	17

Table 7 shows the data elicited by ***Kevin,*** the picture of a young man. This cartoon was not very good at classifying the data sets. The usual chi squared test finds that the classifications were on the whole correct (χ^2=8.7; p=<.05), but a substantial number of cases were classified incorrectly (fully twelve of the L2 cases were classified as L1 speakers, and six of the L1 speakers were classified as L2 speakers. Eight of the L2 speakers were undecided. The distinguishing feature here seems to be that the L1 speakers produced a very low number of singleton responses, and were very likely to produce a response which was also used by L2 speakers.

Table 7: Classified data for Kevin

Table 7a: Examples of words in the corpora for Kevin

Shared responses
SERIOUS SMART UGLY MYSTERIOUS INTELLIGENT EXTROVERT SUSPICIOUS RUDE HAPPY INTERESTED WEIRD JUDGMENTAL SCARY SEXIST SKINNY MISCHIEVOUS WORKER HUNCHBACK ANGRY MANIC COMFORTABLE POOR BORING STARING RURAL FUNNY INTIMIDATING ADULT FARMER WORKING BIG_HEAD MAN ODD SMILING GRUMPY CREEPY SHY SILLY QUIET UNHAPPY

Typical L1 responses
GRITTY CONTENT SMIRKING UNKEMPT DANGEROUS INQUISITIVE FRINGE LAD TOUGH SIMPLE LONELY CONFIDENT HARD STRANGE COMMITTED SLY HIGH_TROUSERS GREASY UNTRUSTWORTHY UNTIDY LOST PAINTER OLDER SARCASTIC DICEY SHADY RUGGED SCRUFFY UNFRIENDLY

Typical L2 responses
CURIOUS INTROVERTED TALL FRIENDLY HAIRY BLONDE THINKING STUBBORN BIG_HEADED PLOTTING LAZY ARROGANT BORED EASY_GOING SELFISH HONEST SHORT WHITE NAIVE GOOD_LOOKING UNTRUSTING STUPID INTERESTING PROUD MIDDLE_AGE ILLITERATE POLITE SAD HANDSOME

Table 7b: Probabilities of a specific response type in the Kevin reference data

	Shared responses	L1 responses	L2 responses	singleton
L1 participants	.423	.236	.050	.291
L2 participants	.277	.050	.264	.409

Table 7c: Discrimination between the 60 response sets for the Kevin reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	16	8	6
L2 speaker response set	12	2	16

The final set of results was generated in response to the child cartoon, ***Cory***. The data is shown in Table 8. This picture was by far the worst of the five cartoons, in that it failed to make clear decisions for more than half the response sets (34 response sets were classified as "undecided"). Only six cases were wrongly classified by Cory: two L1 speaker response sets were incorrectly classified as L2 speakers, and four L2 response sets were identified as L1 speakers. Again, a chi squared analysis suggests that the distribution of the classifications is unlikely to be due to chance (χ^2=8.02; p=<.05), but the overall success rate can only be described as poor.

Table 8. Classified data for Cory

Table 8a: Examples of words in the corpora for Cory

Shared responses
CHEERFUL CREEPY MALE EXCITED LAUGHING SMART NAUGHTY TEENAGER SURPRISED CURIOUS CUTE CHILDISH BOLD IMMATURE OUTGOING ACTIVE ENTHUSIASTIC MISCHIEVOUS VULNERABLE INNOCENT EARS LIVELY EXTROVERT SHORT JOYFUL ENERGETIC FRIENDLY PLAYFUL YOUNG NAIVE HAPPY SMALL CHILD UGLY FUNNY SMILEY CHEEKY

Typical L1 responses
MESSY FRECKLES ANNOYING TROUBLE BOY SPIKY_HAIR BIG_HEADED BIG_FOREHEAD JOLLY INQUISITIVE NUISANCE SPORTY LOUD WILD FUN

Typical L2 responses
SMILING EXTROVERTED EASY_GOING CRAZY SPORT SCARY CHATTY SYMPATHETIC STUBBORN HANDSOME SWEET CHARMING BLUE_EYED OPEN STRANGE GOOFY KIND MEAN NERVOUS SILLY THRILLED LITTLE IMPATIENT

Table 8b: Probabilities of a specific response type in the Cory reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.173	.173	.050	.604
L2 participants	.173	.050	.183	.591

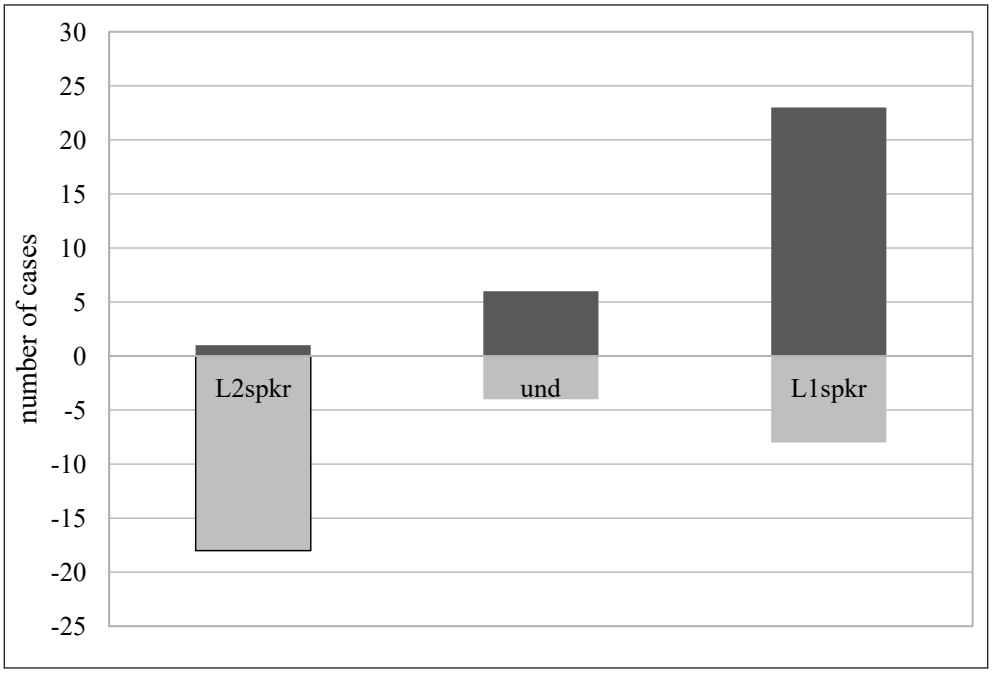
Table 8c: Discrimination between the 60 response sets for the Cory reference file

Actual evaluations	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	10	18	2
L2 speaker response set	4	16	10

4. Discussion

The present study was set to explore how a Bayesian approach could help in distinguishing between NS and advanced English learners from a small set of words provided for the same picture stimuli. According to our results (see Tables 4-8 above), the picture stimuli in the study were actually very inconsistent in how they categorised the response sets. We originally expected that the pictures would tend to categorise an individual test-taker's response sets in the same way, but again this was not the case. None of the sixty test cases was consistently categorised by all five tasks – largely because the Neville and Cory pictures produced relatively large numbers of undecided classifications. If we disregard these two sets of results, we can combine the data from the three best classifiers into a "majority verdict" for each test case. This data is shown in Figure 4. Using this approach, the 60 test cases were generally well-categorised: most response sets were correctly ascribed to the correct group, with only one L1 response set being classified as an L2 example. A small number of response sets were classified as undecided. This distribution is better than chance (χ^2= 25.1; p<.001), but it is clearly not as good as we might have hoped. The main problem here seems to be that a small number of the L2 speakers seemed to generate response sets that were reliably classified as coming from L1 speakers, that is, their choice of words is more characteristic of the choices made by the L1 group. This may not in fact be such a serious problem as we first thought – it may just be a reflection of the very high standard of proficiency enjoyed by some of the L2 participants. If that is the case, then the real problem cases are the instances where the program classifies an L1 participant as an L2 speaker. Figure 4 shows that only 1 L1 speaker was mis-classified in this way when the three best data sets are taken into account. We should also bear in mind that there is always some error rate in estimations, especially if we conceive 'nativelikeness' as a binary category (Vanhove, 2020).

Figure 4: The "majority verdict" from the three good discriminator cartoons (Shirley, Neville and Margaret). L1 response sets are shown in black, L2 response sets are shown in grey



The present study also confirms the idea that small sets of words carry very large amounts of information about L2 speakers' vocabulary use. However, the approach used does not work as well as we expected it to do. With hindsight, it seems that the decision to ask participants to provide only six words in response to the picture stimuli was a tactical error. It resulted in a fairly high number of cases where the analysis was unable to make a confident categorisation. It also allowed the confidence judgments to be strongly influenced by a single instance of an "inappropriate" response. For example, if an L2 speaker generated just one response that was normally generated by L1 speakers, then the confidence estimate would be skewed in the direction of an L1 assessment. If this "inappropriate" word was introduced as the fifth or sixth response, then further evidence would not be available to correct this error. With a larger number of words in the response sets, errors of this sort are normally corrected. Some simulation work with artificially created responses sets suggests that responses sets consisting of 10 words are considerably more powerful than smaller response sets: they almost always result in a definite decision one way or the other. This is a question on the viability of minimal vocabulary tests that should be further explored.

It should be borne in mind that minimal vocabulary tests of this kind are constrained by the task we ask testees to perform. Therefore, the more information we have about how participants approach the task and the type of output it produces in large populations, the better it will be for the interpretation of the scores. In this study we chose these five pictures because we thought that the caricature cartoons would generate a fairly narrow range of responses, especially when we instructed participants to supply us with single-word adjectives. This turned out not to be the case – about 50% of the responses were singleton responses generated by only one participant. More importantly, perhaps, a number of respondents gave us descriptors which focussed on the style of the cartoon, rather than the person who was being depicted. BIG HEADED and BIG EARED both appeared surprisingly often in the response sets. It is not clear whether the same problem would arise if we used other kinds of visuals. Equally surprising was the finding that the cartoons differed quite markedly in the kinds of words that they elicited. We had originally thought that the stylistic similarities between the pictures would result in response sets that were to a large extent comparable, but again this turned out not to be the case. For all five stimulus pictures, the number of singleton responses was considerably larger than we had found in our pilot studies, and consequently, the number of response words that could be classified as typical L1, typical L2 and shared words was correspondingly reduced. Typical L1 words, for example, accounted for only 20% of the L1 speaker responses, and typical L2 responses accounted for only 18% of the L2 speaker responses. Shared responses accounted for around 22% of the responses. However, there was considerable variation around these means: 42% of the responses that L1 speakers made to the *Neville* picture were shared responses, and only 29% of their responses were singletons. And both groups generated about 60% of singleton response for the *Cory* picture. Clearly, there is an issue of stimulus consistency here which needs to be investigated.

It can also be possible that the two groups may have approached the task from a different point of view or used different strategies to provide answers. For example, among the NS responses, we have a number of "awkward" responses (e.g. for Shirley: *singer, teethy*), which feel as though they belong to an informal register, whereas some of the NNS responses seem to be more "literary". When checking the items for frequency and range, we see that words produced more often by learners (1) appear more frequently in the frequency lists, such as the JACET List (Ishikawa et al., 2003) (e.g. *young, shy, thin*), (2) can be more often found in students' textbooks (e.g. *cheerful, talkative, open-minded, friendly*) and (3) can be often cognates (e.g. *elegant, relaxed, attractive, extroverted, modern...*) or borrowings (e.g. *fashion*). NSs sometimes produce words that learners do not typically know or that appear less often in textbooks (e.g. *stern, scruffy, bubbly...*), but we do also find words that tend to be very basic (e.g. *quiet, grandad, boy*). So both groups use a mixture of high and low frequency items, it is not just a matter of frequency or

range: there does not appear to be a reliable significant difference between the groups in respect of these features. Register does seem to be an important feature, and some individual responses are strongly marked for this. However, we do not find that the majority of the responses generated by a single individual are characterised in this way.

We also think that the study raises some interesting questions about the use of a Bayesian approach to linguistic data of this kind. We have observed there are some technical issues that will need to be resolved in the future. In this study, we started out with two collections of response sets each generated by 80 test-takers. Each collection was split into two: fifty responses sets were used to define a reference corpus, and thirty response sets were held back to be used as test cases. Of course, these two numbers are arbitrary: we could have split the data in other ways. For example, we could have used a set of 25 response sets to establish the reference corpus, and this would have left us with 55 response sets to be used for evaluation. Or we could have used a bigger number of response sets to establish the reference corpus, and evaluated only a handful of test cases. Ideally, we would like to work with a small but reliable reference corpus, since this makes it considerably easier to build the reference corpus, and allows us to evaluate a larger number of test cases. Unfortunately, we do not know how the size of the reference corpus affects the evaluations. We might expect that increasing the number of response sets that are used to build the reference corpus would increase the number of singleton responses, but exactly how this interaction would work is unclear. We might expect that a larger reference corpus would affect the number of response words that are "typically" L1 responses or "typically" L2 responses, but it is difficult to assess this characteristic in practice. We are currently assembling some very large data sets which will allow us to answer these questions with some confidence (see Meara & Miralpeix, in preparation).

A more important issue concerns the way we have characterised the four types of words in the response sets, particularly the singleton responses. In this paper, we have treated any word which appeared only once in the relevant reference corpus as a "singleton", and we have lumped together into a single class words that were generated by a single L1 speaker, or a single L2 speaker. Any new word that appeared in the test response sets, but not in the reference corpus was classified as a singleton, regardless of its characteristics. It is probable that this classification is just too broad, and that a closer examination of the singletons generated by the L1 group and by the L2 group would reveal some subtle differences between the groups. For example, the L1 singletons tend to be infrequent words, whereas the L2 singletons are sometimes invented words based on cognates. We have not explored this avenue here, as it is difficult to automate the process of distinguishing the different types of singletons. However, a closer examination of these words would be worthwhile. Simply ignoring the singleton problem, and treating words that appear in the reference corpus as L1

words or L2 words regardless of how many times they occur would be an even simpler solution. This approach would have the added advantage of increasing the proportion of "typical" L1 and "typical" L2 words, but again, it is not clear how this approach would affect the performance of the program.

A related issue has to do with our criterion for classifying a word as a "typical L1 word" or a "typical L2 word". In this study, all words which occur at least two times but only in response sets generated by L1 speakers were identified as L1 words, and all words which occur at least two times but only in response sets generated by the L2 speakers were identified as L2 words. However, once again, we are dealing with an arbitrary cut-off here. We could have used a rather stricter criterion, in which case the number of words identified as "typical" cases would have been much smaller, and we would need to introduce a new category of "words which do not occur often" - say, all words which occur only once or twice in the reference corpus. We think that this would make the classifier program rather less accurate than it is currently. Alternatively, we could lower the threshold for describing a word as "typical", and include all words which are generated by only one of the groups. This would increase the number of "typical" words, and would allow us to eliminate the entire class of singleton words by subsuming them into the "typical L1 word" and "typical L2 word" categories. Our guess is that this might be a good way to go in future research of this kind.

The last technical issue concerns a feature that we have not commented on before, but will doubtless have been noted by astute readers. Given that the reference corpora used in this study are samples, and not comprehensive lists, there will always be occasions when, for example, an L2 speaker uses a word which has formally been defined as a "typical L1 word" because it has not appeared as a "typical L2 word" in the reference corpus. The question which arises here is how should we deal with these cases. The logical solution would be to say that, by definition, L2 speakers do not use "typical L1 words", and therefore the probability of an L2 speakers using a word of this type is nil. The problem with this obvious solution is that with these assumptions, and working through the steps in Table 3, a L2 response set that contains a single instance of a "typical L1 word" will return a value of zero despite the fact that the response set was actually generated by an L2 speaker. And once this zero value is found, it cannot be changed by any later data because of the way the mathematics works. Obviously, we need to avoid this over-determination, and we do this by setting a non-zero value to the probability that an L1 word will be generated by an L2 speaker and vice-versa. In tables 4-7 we have set these values to 0.05 – i.e. we anticipate that an L2 speaker might produce a "typical L1 response" from time to time: usually, slightly fewer than one response of this type per response set. This value of 0.05 is actually quite strict, and it severely penalises a test-taker who produces "the wrong sort of word". Ideally, we would like this non-null parameter to be an empirically based one, rather than

an arbitrary choice. As usual, we do not know how changing the non-null parameter from 0.05 to a rather higher figure would affect the way the classification program works. We think it should result in fewer incorrect classifications, but that it might generate more undecided classifications. This is an issue that we can address using the simulation approach mentioned earlier.

5. Conclusion

To sum up, this paper presents an empirical way of measuring productive vocabulary. Data from a minimal vocabulary test taken by NS and advanced EFL learners was analysed following a Bayesian approach, which was used to decide whether the data was generated by an L1 or an advanced L2 speaker. In theory, it seemed a good way to put this method to the test, as at high proficiency levels the differences in lexis between learners and native speakers may not be obvious (Hellman, 2008), and even less in this case with sets of just six words. Therefore, results from the current study can inform future research on the suitability of this method to distinguish between learners at different proficiency levels, where differences in productive vocabulary are more remarkable. This form of assessment could also be very helpful for teachers: by obtaining this information from minimal vocabulary tests, they could more easily identify students' weaknesses in vocabulary skills. We think the format might be particularly useful in situations where testing events need to be administered frequently, as administration of the test in its current format requires only a very short time. This is a considerable advantage over more traditional vocabulary tests. Despite this, the data the test provides appears to be rich, and the test format is challenging, even for advanced test-takers.

Finally, the work we have reported here has turned out not to be as straightforward as we expected and we have identified a number of technical issues that we failed to anticipate. In spite of this, we think the idea of assessing productive vocabularies using minimal vocabulary tests and Bayesian statistics might be worth of further exploration. In particular, we can speculate whether the Bayesian probabilities generated by the program correlate with scores generated by the productive vocabulary tests that we discussed in our introduction. Work of this sort clearly lies outside the scope of this paper, but we think that it would be worth doing work of this kind in future.

Acknowledgements

The authors would like to thank Dr Rhian Meara for her help in data collection, as well as all the students who participated in the study. Thanks also to the editor and reviewers of VIAL for their comments on the article.

6. References

Castañeda-Jiménez, G., & Jarvis, S. (2014). Exploring lexical diversity in second language Spanish. In K.L. Geeslin (Ed.), *The Handbook of Spanish Second Language Acquisition* (pp.498-513). Chichester: Wiley Blackwell.

Coxhead, A., Nation, P., & Sim, D. (2014). Creating and trialling six forms of the Vocabulary Size Test. *TESOLANZ Journal*, 22, 13-26.

Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly,* 51(4), 844-867.

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

Hellman, A. (2008). *The limits of eventual lexical attainment in adult-onset second language acquisition*. PhD thesis, School of Education, Boston University.

Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N. & Tono, Y. (2003). *JACET 8000: JACET List of 8000 basic words*. Tokyo: JACET.

Jiménez Catalán, R.M. (2014). *Lexical Availability in English and Spanish as a Second Language.* Berlin: Springer.

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

Laufer, B., & Nation, I.S.P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-323.

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

McGrayne, S.B. (2011). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines & emerged triumphant from two centuries of controversy*. Boston MASS.: Yale University Press.

Meara, P.M., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142-154.

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

Meara, P.M., & Miralpeix, I. (2015a). *V_YesNo.* Cardiff: Lognostics.

Meara, P.M., & Miralpeix, I. (2015b). *V_Size.* Cardiff: Lognostics.

Meara, P.M., & Miralpeix, I. (2017). *Tools for Researching Vocabulary*. Bristol: Multilingual Matters.

Meara, P.M., & Miralpeix, I. (in prep.). *Minimal vocabulary tests for English language learners.*

Meara, P.M., & Olmos Alcoy, J.C. (2010). Words as species: An alternative approach to estimating productive vocabulary size. *Reading in a Foreign Language*, *22*(1), 222-236.

Melka Teichroew, F.J. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp.84-102). Cambridge: CUP.

Miralpeix, I. (2020). L1 and L2 vocabulary size and growth. In Webb, S. (Ed.). *The Routledge Handbook of Vocabulary Studies* (pp. 189-206). New York: Routledge.

Nation, I.S.P. (1984). *Vocabulary Lists: Words, affixes and stems*. Victoria University of Wellington: English Language Institute. Occasional publications 12.

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

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(1), 9-13.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning 68*(4), 1032-1075.

Pearl, L., & Goldwater, S. (2016). Statistical learning, inductive bias, and Bayesian inference in language acquisition. In J.L. Lidz, W. Snyder, & J. Pater (Eds.), *The Oxford Handbook of Developmental Linguistics* (pp.664-695). Oxford: OUP.

Roghani, S., & Milton, J. (2017). Using category generation tasks to estimate productive vocabulary size in a foreign language. *TESOL International Journal*, *12*(1), 128-142.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88.

Stone, J.V. (2013). *Bayes' Rule: A tutorial introduction to Bayesian Analysis*. Sheffield: Sebtel Press.

Vanhove, J. (2020). When labelling L2 users as nativelike or not, consider classification errors. *Second Language Research, 36*(4), 709-724.

Webb, S. (2018). Receptive and productive vocabulary size of L2 learners. *Studies in Second Language Acquisition, 30*(1), 79-85.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics, 168*(1), 34-70.

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

Xue, G., & Nation, I.S.P. (1984). A University Word List. *Language Learning and Communication* 3 (2), 215-229.

Zinszer, B.D., Rolotti, S.V., Li, F., & Li, P. (2018). Bayesian word learning in multiple language environments. *Cognitive Science* (42, Suppl.2), 439-462.