_lognostics
tools for vocabulary research

A methodology for evaluating the effectiveness of vocabulary treatments
Paul Meara and Ignacio Rodríguez Sánchez ***Swansea University***

1. Introduction

A persistent and awkward problem in vocabulary acquisition research arises when we want to evaluate the effectiveness of different ways of teaching vocabulary. This problem is very simple to state. Ideally, when we have a question of this sort, we construct an experiment in which all the relevant variables are controlled, and the effects due to the different methods are isolated. In real life, of course, it is very difficult to do this where vocabulary acquisition is concerned. For obvious reasons, simply teaching a set of words using method A and the same set of words using method B to a single set of learners does not work: once learners have learned the target words with one method, they cannot be asked to learn the same words again using different method. Most people get round this problem in one of two ways. Either they teach two different sets of words to a single group of testees, or they teach a single set of words to two different groups of testees. Neither of these solutions is completely satisfactory, however. In the first case, we have the problem that it is difficult to draw up two sets of words that are truly equivalent when we do not know for certain what factors affect the learnability of words. In the second case, we have the problem of ensuring that our two experimental groups are truly equivalent: although we can control for gross effects, it is not easy to set up large groups of testees who are all ignorant of exactly the same set of target words. Clearly, then, it would be very useful if we could develop a methodology which allowed us to examine the effectiveness of vocabulary teaching treatments, and at the same time allowed us to sidestep these awkward problems. This chapter is a preliminary exploration of a methodology that might allow us to do this. The next section explains the basic methodology, and the third section reviews some new data which tests how well the theoretical model advanced below actually works in practice.

2. The theory

The model that we have been developing is based on three very simple assumptions.

Firstly, we assume that vocabulary acquisition can be described in terms of a model which has a small number of states. For a given learner, and for a given set of target words, we assume that it is possible to describe each word as being in one of these states. For the purposes of illustration, let us assume that we are dealing with a three state model. Words which are not known by the testee are deemed to be in state one; words which the testee 'knows' are deemed to be in state three; words which the testee is unsure about are deemed to be in state 2. (The choice of three states here is an arbitrary one and it is possible to expand model by differentiating state 2 more finely so the different types of uncertainty are separated out. An example of this will be found in the third section, where we use a four state model. For the purposes of explanation, however, we will stick with a three state model, as it makes the exposition clearer and simpler.) The important thing to note about this model is that, unlike some of the implicit models described in the L2 literature (e.g. Palmberg 1987; Ringbom 1987), it is NOT a linear model. In our model, there is no 'cline' or 'continuum' for words to move along gradually. At any given time, all words in the target set fall unambiguously into one of the three states, but they are free to move into any of the other states without hindrance, should the circumstances allow.

The second assumption that we make is that we can sensibly ask learners to report on how well they know words. Obviously, there are problems with this assumption. Some learners deliberately lie about

their word knowledge; some learners think that they know a word when in fact they are confusing it with a different one; some learners claim to know a word when they have little more than a nodding acquaintance with it, while others will admit to knowing a word only if they are 100 percent certain what it means and how to use it. Clearly, too, there are problems with the whole concept of 'knowing a word' (cf. Richards 1976; Nation 1990). Nevertheless, we think it is possible to ignore these problems for the time being. Accepting that IS possible to get meaningful self-assessments from learners about their vocabulary knowledge allows as to take our model a stage further. We do this by showing Ss a large number of target words, one at a time, and asking them to assign these words into one of a number of categories. For instance, we can ask testees to assign all words they are sure they know to category 3, and words they are sure they do not know to category 1. Judgments of this sort appear to be easy to make, and this allows as to collect large amounts of data very easily. Using a simple computer program, for example, allows as to collect hundreds of judgments of this sort in one 10 to 15 minute session. At the end of the testing session, we can describe the testee's knowledge of the target vocabulary in terms of a simple vector, such as the one shown in the T1 column of Table 1. This table shows that at Time 1 our (fictional) testee claims to know 50 words out of a target set of 300, to be unsure about 50 other target words, and not to know the remaining 200 words.

Table 1. Two illustrative vocabulary vectors for one (imaginary) testee.

	Test T1	**Test T2**
State 1 – I don't know this wd	200	117
State 2 – not sure	50	110
State 3 – I do know this wd	50	73

Our third assumption is that the data we get from experimental testing sessions of this sort is not entirely stable. Our experience is that repeated testing of the same words rarely produces identical results, especially when the words being tested are on the edge of the testees' confidence. We think that this reflects the fact that vocabularies are not static: words known one-day can be forgotten the next, while words not known today may be easy to bring to mind tomorrow. In this model, vocabulary is in a constant state of flux, with words moving between three states relatively freely. This means that if we test our imaginary testess again after, say, a week's delay, we might find that the same target words produce slightly different results, as shown in the T2 column in table 1. This data suggests that at T2 our testee 'knows' 73 of the target words, and still fails to know 117 target words, while he still unsure of 110 words. The obvious explanation of this is 83 words moved from State 1 to another state: 60 to State 2, and 23 to State 3. In reality, this is not likely to be the case. Much more likely is that there was some movement between all three categories. This suggests that we need a rather more sophisticated analysis of our data, and an analysis of this sort is shown in Table 2. The left-hand part of Table 2 shows that of the 200 words that were in State 1 at T1, 100 remained in that state, 60 moved up to State 2, while 40 move to State 3. Conversely, of the 117 words that appear in State 1 at T1, 100 and were also in that state at T2, seven were in State 2 at T2, and ten were in State 3 at T2.. The right-hand part of table 2 shows the same data in a slightly different format. Here, the figures show not the absolute number of words that fall into each of the categories, but rather the probability of a word in any specified state moving to another state at T2. The probability figures provide us with a **transitional probability matrix (TPM)**. It allows as to examine the patterns of movement in the vocabulary of our (fictional) testee without reference to the actual numbers of words tested. The matrix shows, for example, that in this data sets, any word has at best a 50% chance remaining in the same state from T1 to T2, and that words are most likely to move towards state two.

Table 2. Raw data and a Transitional Probability Matrix for one (imaginary) testee.

	T2	T2	T2		T2	T2	T2
	S1	**S2**	**S3**		**S1**	**S2**	**S3**
T1 State 1	100	60	40		.50	.30	.20
T1 State 2	7	25	18		.14	.50	.36
T1 State 3	10	25	15		.20	.50	.30

To summarise, deriving a TPM for a particular testee is simply a question of collecting data on two separate occasions, and calculating the probability of movement between states. However, once this initial TPM has been constructed, we have in our hands are very powerful tool, which allows as to make long-term forecasts about what a testee's vocabulary will look like in the future. The probabilities in Table 2 record the actual movement of words between the three states from T1 to T2. If we now assume that these probabilities are reasonably stable and do not change much over time, then we can use the T1/T2 to matrix to forecast how many words will be in each of the designated states at a later testing time, T3. We do this by multiplying the T2 vector by the T1/T2 matrix. This calculation is carried out as follows.

Table 2 tells us that of the 117 words in State 1 at T2, 50% will remain in State1 at T3, 30% will move to State 2 and 20% will move to State three. This gives us:

state 1	**state 2**	**state 3**
117*.50=59	117*.30=35	117*.20=23

Similarly, the matrix tells us that of the 110 words in State 2, 14 % will move to State 1, 50% will remain in State 2, while 36 percent will move to State 3. This gives us:

state 1	**state 2**	**state 3**
110*.14=15	110*.50=55	110*.36=40

And for the words in State 3, the matrix tells us that 20% will move to state 1, 50% will move to State 1 and 30% will remain in State 2. This gives us:

state 1	**state 2**	**state 3**
73*.20=14	73*.50=37	73*.30=22

Adding these figures together gives us a new distribution for the target words at T3, and this vector is shown in the T3 column of Table 3.

We can now apply the same process in an iterative fashion, multiplying the T3 vector by the matrix to derive a T4 vector, multiplying the T4 vector by the matrix to produce a T5 vector and so on. The results of repeating this iteration process nine times are shown in Table 3. The interesting thing to note here is that iterating this multiplication procedure over a number of intervals eventually produces a stable pattern, a vector which does not change when it is multiplied by the matrix. This 'eigen vector' is not to be taken as showing that the vocabulary has fossilised, or that there is no longer any movement between states. The probability of words moving between the states remains the same as it was in the

Table 3. Repeated multiplication of a vector by a matrix

	T1	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*T8*	*T9*
State1	200	117	88	79	76	75	74	74	74
State2	50	110	127	132	134	135	135	135	135
State3	50	73	85	89	90	90	91	91	91

early iterations of the model: only 50% of words in State 3 still remain in that state at each iteration, but the 50% of words that move out of State 3 are exactly balanced by other words moving in. This gives an illusion of stability in the overall numbers.

The emergence of an eigen vector is a general property of processes which can be described in matrix terms (see Bradley and Meek 1986 for an accessible discussion of the principles). The eigen vector seems to represent an equilibrium state to which a system would naturally evolve if it were allowed to do so. Meara (1989) suggested that in the case of vocabulary acquisition, it might be possible to use eigen vectors as a long-term forecast of the way a student's vocabulary might develop over a long period of time if the initial learning conditions persist. We have used this idea very effectively to make uncannily accurate forecasts of the number of words a testee will know after repeated readings of a long text, for example (Horst and Meara 1999). The question we ask here is whether this method can be generalised more widely to 'naturalistic' vocabulary acquisition.

3. Experiment

This section evaluates the claim that it might be possible to forecast the long-term vocabulary uptake of a group of learners using the matrix models described in the previous section. We do this by assessing the vocabulary knowledge of a group of advanced learners of Spanish at three times times, T1, T2 and TX. T1 and T2 are close together, and we use the data from these two tests to make a long-term forecast about the testees' overall vocabulary knowledge some weeks later. We then compare these forecasts with the actual data obtained at TX. The claim we are interested in testing is whether there will be a close correlation between the forecast scores and the actual scores at TX.

3.1 subjects

A group of 28 undergraduates following courses at Swansea University took part in the study. All were native speakers of English, and participated on a voluntary basis. Six data sets were lost for technical reasons, leaving a complete set of 22 data sets in the experimental pool.

3.2 materials

A list of 360 words was developed, comprising 180 words from frequency bands six and seven and 180 words from frequency bands nine and ten of the Juilland and Chang-Rodriguez frequency list (Juilland and Chang-Rodriguez 1964). These words were selected on the grounds and they were likely to be met by students at this level during the normal course instruction, and were words that an advanced non-native speaker might be expected to know: i.e. though all the words were not very high frequency items, they were definitely not arcane or recherché words. (J&C-R's levels six and seven correspond to rank orders between 3500 and 4000 words; levels six and seven correspond rank orders 4500 to 5000 words.) These words were assembled into a single test, and an extract from this text is shown in Table 4.

Table 4. Extract from the test instrument.

1: palma ❑	**2: cuna** ❑	**3: águila** ❑	**4: sazón** ❑
5: homenaje ❑	**6: almanaque** ❑	**7: galán** ❑	**8: pesadumbre** ❑
9: abrumado ❑	**10: sede** ❑	**11: garantizar** ❑	**12: traba** ❑
13: esparto ❑	**14: caudillo** ❑	**15: almacén** ❑	**16: resbalar** ❑
17: vicisitud ❑	**18: traición** ❑	**19: exaltado** ❑	**20: muelle** ❑

.....

Testees were asked to rate each of the 360 words on a four point scale, with the points labelled as follows:

1. I think I have never seen his word before, and I do not know what it means;
2. I have seen this word before, but I definitely do not know what it means;
3. I have seen his word before, but I'm not sure what it means;
4. I have seen his word before and I am sure I know what it means.

The tests were administered three times. T1 and T2 were separated by a period of three weeks. The same items we used in both test occasions but the order of the items was varied as a way of minimising practice effects. TX, which also used the same words, was administered approximately 15 weeks after T2. Each test session lasted approximately 20 minutes. The testees were not exposed to any special form of instruction during the period of study. They simply followed their normal classes throughout.

3.3 Analysis

For each testee, we constructed a 4x4 matrix using the data from T1 and T2, capturing the way words moved between the four states from T1 to T2. The matrices were then used to generate a long-term forecast for each testee, using the method described in the previous section. Briefly, we obtain a vocabulary vector that describes the state of the 360 target words at T2, we then multiply this vector by the matrix to obtain a new vector, which in turn is multiplied by the matrix to generate a further vector. This iterative process is repeated until the input and output vectors are identical. We take this eigen vector to be the long-term forecast for that particular testee.

Data for all 22 testees are summarised in Table 5 and Figure 1 below. The data reproduced here is a simplification of the full data set, in that it deals only with category four words -- words that Ss claimed to be absolutely sure they knew.

Table 5: Mean words in State 4: actual and forecast

	Actual	Forecast
Mean	174.9	186.09
sd	80.78	100.81

Table 5 shows the mean number of words that each testee assigned to State 4 at TX. The table also shows the mean forecast number of State 4 words. The actual data and the forecasts are almost identical, and the difference between the two data sets is nowhere near significant (t=1.33, p=.198).

Figure 1: Number of items in State 4: actual vs forecast

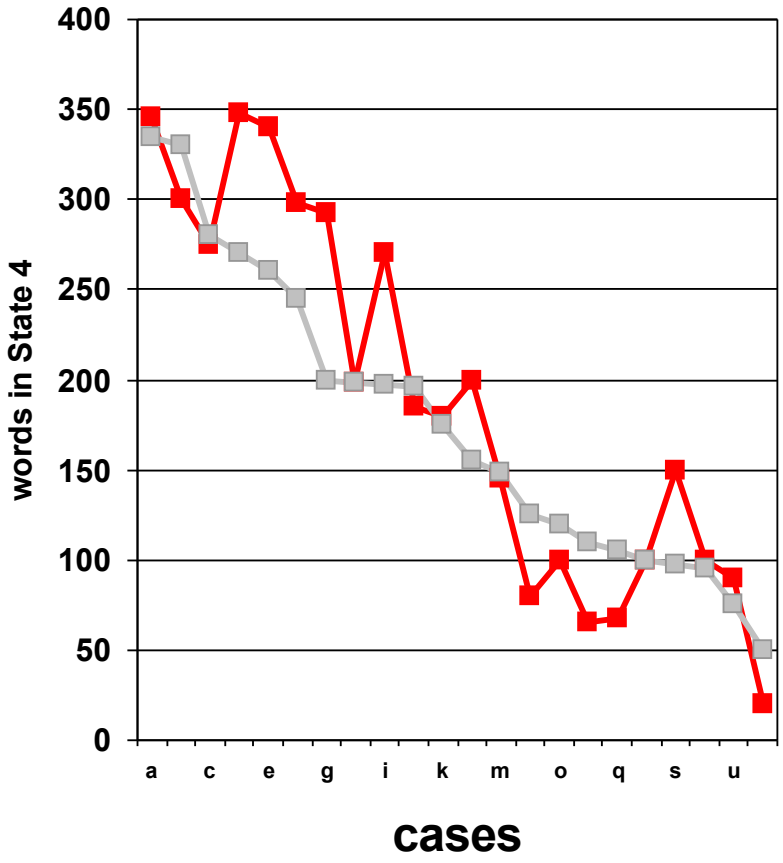


Figure 1 shows the same data disaggregated for individual testees. In Figure 1, the dark line represents the forecast values for the TX test and the lighter line represents the actual value for each testee. Testees are ranked from left to right according to their actual scores on the TX test. The data suggest that there is a remarkably high degree of fit between the matrix forecasts and the reported data. The correlation between the two data sets is .933. This is a highly significant relationship (p<.001) and strongly supports the claim that we are able to make reasonably good forecasts about long-term

vocabulary uptake in testees at this level. Figure 1 suggests that we may be slightly overestimating the true scores of testees with larger vocabularies, and slightly underestimating the true scores of testees with smaller vocabularies. Cases where the matrix forecast is much higher than the actual data tend to cluster on the right-hand side of this figure (cases n,p,q and v), while cases where the matrix forecasts underestimate the actual score tend to cluster at the left the figure (cases d,e,f,g and i.) This is a pattern of results that we have observed in other studies that we have carried out, though at the moment we do not have a good explanation as to why the matrix forecasts are biased in this way.

4. Discussion

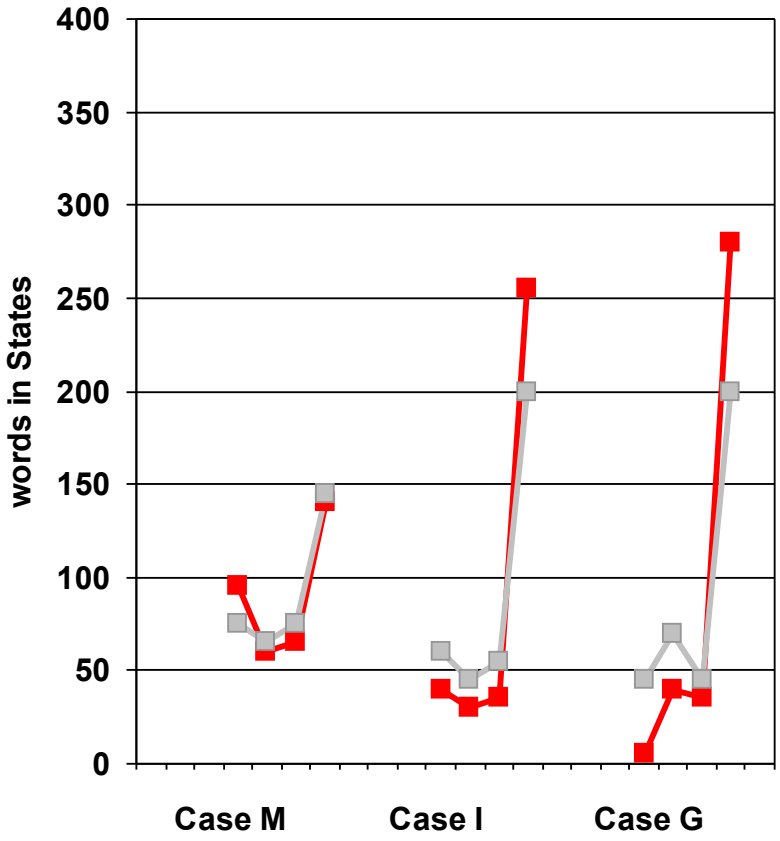
In the introduction to this chapter, we observed that it was difficult to evaluate claims about vocabulary instruction because practical problems forced us to adopt compromises in our experimental designs. The data reported here suggest that we might be able to avoid these awkward compromises by using matrix-based, long-term forecasts about vocabulary uptake. If we could do this, then we would be in a position to devise experimental evaluations in which each testee's scores in an experimental treatment are compared against their own forecast scores in the absence of a treatment. This turns a between-group problem into what is effectively a repeated-measures problem, reducing the need for large groups of testees, and removing the problem of finding equivalent word lists for different treatments. Basically, we can measure the effectiveness of a vocabulary treatment in the following way. First, we make a long-term forecast of how much vocabulary we would expect students to know under normal circumstances -- that is without the treatment. Next, we introduce the treatment and let it run for a reasonable length of time. Finally we compare actual vocabulary uptake at the end of the treatment period with the long-term forecasts. If treatment produces significantly better results than the forecasts, then we can conclude that the experimental treatment is an effective one.

There are, in fact, some grounds for thinking that this conclusion may actually be a cautious one, and that the matrix model may actually be better than it appears to be here. The main reason for this suggestion concerns the timing of the three tests. We suggested earlier that the eigen vector values typically emerge after seven or eight iterations of the multiplication procedure. Since the time difference between T1 and T2l in our study is three weeks, seven to eight iterations would be equivalent to 21 to 24 weeks' study time. This is considerable longer than the time available to us for our own study. The real data summarised in figure 1 was collected after only 15 weeks – i.e. the equivalent of five iterations in the matrix. The implication of this is that the real data may have some way to go before it reaches the equilibrium levels indicated by the forecasts. Had we had more time at our disposal, we would have been able to test this conjecture more adequately. As it is, we have the clear possibility that we could improve on the very high correlation is already reported.

A second reason for suggesting that our forecasts may be better than we think emerges from a detailed consideration of some of the individual cases in the testee group. In Figure 2, we show three examples of the distribution of 360 target words in each of the four states -- that is the number of words State 1, in State 2, in State 3, and State 4, -- after test TX, and compare this distribution with the forecast distribution for three individual Ss. Case M shows the full data set for Testee M, one of the cases where the matrix model makes a very close set of forecasts, while Case I shows the complete data set from Testee I, a case with a very large discrepancy between the actual data and the matrix forecast. The forecast distribution for Case M is very accurate, while even the relatively inaccurate forecast in Case I still preserves the overall shape of the actual distribution, correctly reporting the rank order of the states, if not the exact values.

Figure 2 also shows data from the case with the largest discrepancy between the actual and the forecast score, Cases G, one of a few cases in the entire data set where the forecasts and the actual data are

Figure 2: number of items in all four states: three case analyses



completely at odds with each other. Interestingly, case G is unusual in that her score on test TX was actually lower than her score on both of the preliminary tests. This is clearly an anomalous result, implying as it does an overall loss of vocabulary over the entire test period. It is not entirely surprising, then, that the matrix model fails to handle the case of this sort satisfactorily.

The interesting thing here, however, is that when then matrix model fails it generally does so by over-estimating the number of words that will appear in state 4. In the context of this chapter, this is important. It means that the long-term forecasts generated by the matrix model will be generous ones, which tend, if anything, to overestimate the vocabulary uptake of learners under normal conditions. Consequently, when we come to compare actual vocabulary uptake under experimental conditions against these forecasts, it is that much less likely that spurious significant differences will emerge, and we can be that much more confident that any significant treatment effects we identify are actually genuine.

5. Conclusions

The methodology that we have outlined here is in reality not much more than a first step in an interesting direction. Clearly, much work remains to be done with matrix models before we can rely on them to provide surrogate data of the kind we need. We believe, however, that the approach we have outlined here might turn out to be a powerful tool in vocabulary research. Specifically, it might allow

us to move beyond tightly controlled studies were small numbers of testees learn small number of words in highly constrained laboratory tasks. This is not to deny the value of laboratory studies, of course: they are important for their rigour, and the way they focus our attention on detail. Nonetheless, we do urgently need to find ways which will allow us to extend this rigour beyond detail and into larger studies. In the 1990s, some widely quoted and influential laboratory research on vocabulary acquisition emerged from the Groningen applied linguistics group while it was under Arthur van Essen's leadership (Mondria and Wit-de Boer 19910. We very much hope that the ideas containing this chapter will contribute to the vigorous debate that arose from that work, and allow it to be extended from the laboratory into large-scale, real-world vocabulary learning tasks.

References

Bradley, I and RL Meek
Matrices and Society. Harmondsworth. Penguin. 1986.
Horst, M and PM Meara.
Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review* 56,2(1999), 308-328.
Juilland, AG and E Chang-Rodríguez
Frequency Dictionary of Spanish Words. Berlin: De Gruyter. 1964.
Meara, PM
Matrix models of vocabulary acquisition. *AILA Review* 6(1989), 66-74.
Mondria, J-A and M Wit-de Boer
The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics* 12(1991), 249-267.
Nation, ISP
Teaching and Learning Vocabulary. Boston: Heinle and Heinle. 1990.
Palmberg, R
Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition* 9(1987), 201-220.
Richards, JC
The role of vocabulary teaching. *TESOL Quarterly* 10(1976) 77-89.
Ringbom, H
The Role of the First Language in Foreign Language Learning. Clevedon: Multilingual Matters. 1987.

Notes

This paper first appeared in: In: **M Bax, and J-W Zwart** (Eds.) *Reflections on Language and Language Learning*. Amsterdam: Benjamins. 2001. 267-278.