Lexical Frequency Profiles:
A Monte Carlo Analysis

PAUL MEARA
University of Wales Swansea

This paper reports a set of Monte Carlo simulations designed to evaluate the main claims made by Laufer and Nation about the Lexical Frequency Profile (LFP). Laufer and Nation claim that the LFP is a sensitive and reliable tool for assessing productive vocabulary in L2 speakers, and they suggest it might have a serious role to play in diagnostic evaluations of learners. The simulations suggest that LFP is not in fact all that sensitive. It works best when the groups being compared have very disparate vocabulary sizes, and is probably not sensitive enough to pick up modest changes in vocabulary size.

1. LEXICAL FREQUENCY PROFILES

This paper is a critical analysis of the Lexical Frequency Profile (LFP) approach to estimating productive vocabulary size described in Laufer and Nation (1995). LFP was originally developed by Nation as a way of assessing whether a particular text is suitable for use with learners at a specified level of proficiency. In its simplest form, LFP takes a text as raw input, and outputs a profile that describes the lexical content of the text in terms of frequency bands. In Nation's original formulation of LFP (Nation and Heatley 1996), the bands are described as follows:

The first [band] includes the most frequent 1000 words of English. The second [band] includes the 2nd 1000 most frequent words, and the third [band] includes words not in the first 2000 words of English but which are frequent in upper secondary school and university texts from a wide range of subjects. All of these base lists include the base forms of words and derived forms.

This use of frequency bands to characterize vocabulary is a fairly standard practice in L2 vocabulary studies. There is some disagreement among scholars about which frequency list provides the best standard, but this is really only a serious issue at low levels of frequency. The standard frequency counts are in broad agreement about which words should appear in a ‘first thousand’ or a ‘second thousand’ word list. It might be argued that the words appearing in lists of this sort would be very dependent on context or genre, but in practice this appears not to be the case. LFP in its current incarnation
uses Nation’s (1986) word lists. These lists have become something of a standard reference in L2 vocabulary studies, and provide good coverage of the genre that Laufer and Nation work with: short texts generated by learners of English.

Typically, output from LFP looks something like Table 1, which shows the number and percentage of word types and word tokens from the text being analysed. In this particular passage—in fact, the first two paragraphs of this paper, with the proper names removed—the majority of words come from the first 1k band (75 per cent of tokens), a small number of words come from the 2k band (10.5 per cent of tokens), and a smaller number of words still (6.3 per cent of the tokens) are taken from the UWL list—a list of words that are objectively low in frequency, but are important in the context of Academic English (Coxhead 2000). Finally, a handful of items (8 per cent of tokens) are not included in these lists, and are counted as NiL (= not in the lists) words. Intuitively, this approach to evaluating texts seems to be very satisfactory, and a number of sources (e.g. Valcourt and Wells 1999; Nation and Waring 1997) present evidence to show that the approach works well in practice.

Laufer and Nation (1995), however, take this idea a stage further. In that paper, they argue that it might be possible to use LFP to gauge the extent of the productive vocabulary available to non-native speakers of English, an idea which Nation (2001: 362) pursues further. Laufer and Nation (1995) argue that the texts produced by learners at different levels of proficiency have characteristically different profiles, and that the text profiles change in predictable ways as learners’ vocabularies grow. This is an attractive idea for vocabulary research, which is seriously short of good tools to assess productive vocabulary abilities, and the claims that L&N make for LFP suggest that it needs to be taken very seriously indeed. These claims are that LFP:

- is a reliable and valid measure of lexical use in writing;
- provides similar stable results for two pieces of writing by one person;
- discriminates between learners at different proficiency levels;
- correlates with an independent measure of vocabulary knowledge;
- is a useful diagnostic test;
- is a sensitive research tool (Laufer and Nation 1995).

<table>
<thead>
<tr>
<th>Word list</th>
<th>Tokens (%)</th>
<th>Types (%)</th>
<th>Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>216 (75.3)</td>
<td>101 (71.1)</td>
<td>84</td>
</tr>
<tr>
<td>Two</td>
<td>30 (10.5)</td>
<td>14 (9.9)</td>
<td>10</td>
</tr>
<tr>
<td>Three</td>
<td>18 (6.3)</td>
<td>15 (10.6)</td>
<td>14</td>
</tr>
<tr>
<td>Not in the lists</td>
<td>23 (8.0)</td>
<td>12 (8.5)</td>
<td>?</td>
</tr>
<tr>
<td>Total</td>
<td>287</td>
<td>142</td>
<td>108</td>
</tr>
</tbody>
</table>

Table 1: VocabProfile analysis of an illustrative text
Clearly, a test with all these characteristics is one that we would expect to be taken up quickly by researchers, and this has indeed been the case. Apart from Laufer and Nation themselves (e.g. Laufer 1994; Nation 2001), LFP has been utilized by a number of other authors, including Muncie (2002), Cobb and Horst (2002), and Meara, Lightbown, and Halter (1997).

My own feeling, based on some work using an LFP-type analysis of radio broadcast texts (Meara 1993), analysis of teacher talk (Meara, Lightbown, and Halter 1997), and texts produced by learners (Bell 2003) is that LFP is more problematical than Laufer and Nation’s (1995) report and some later accounts of this work suggests. There are several reasons for this. First, there are some technical problems and uncertainties with the way the texts are processed. These concern the way errors are handled (do we ignore errors or correct them?), the way proper nouns are counted (do we count proper names as low frequency items simply because they are not in the word lists?), the treatment of formulaic sequences (do we treat ‘Victoria Park’ as two separate words, or as a single proper noun?), and so on. Laufer and Nation are not completely clear on these points. Secondly, there are also some problems with the way the profiles are reported. Laufer and Nation’s original paper reports a series of analyses comparing the performance of their subject groups on each point of the profile, but this ignores the fact that the profile scores are made up of interdependent scores: if 80 per cent of the words I use in a text come from the 1k frequency band, then the highest score I can get on the 2k band is 20 per cent, and my scores on the other parts of the profile are constrained in a similar way. For this reason, it does not make sense to analyse each frequency band separately. Finally, and more importantly, in my experience, texts do not vary as much as Laufer and Nation’s analysis seems to imply they ought to. In most of the texts I have worked with, the proportion of 1k words varies by a few percentage points around 80 per cent, and the percentage of UWL and NiL words varies by a few percentage points around 10 per cent. Normally, it is difficult to get texts of more than a couple of hundred words from learners in a typical testing situation, and this means that ‘a few percentage points’ will in practice translate into ‘a few words’. It is difficult to see how the occurrence or non-occurrence of just four or five words could possibly carry the sort of informational burden that Laufer and Nation are asking them to carry.

This suggests that LFP really needs to be subjected to an in-depth regime of evaluation before we allow it to become a standard analytical tool. It is, however, very difficult to test claims of the sort Laufer and Nation make by running experiments with large groups of learners, simply because of the complexity of the logistics. Suppose, for example, that we want to compare the profiles of texts generated by learners with vocabularies of, say 4,000 and 6,000 words respectively, and that we want to show that these profiles differ systematically. It is very difficult to set up large groups of learners with these precise characteristics, and even if we could set up the groups, it is by no means a simple task to collect large quantities of data of the sort reported in
Laufer and Nation’s original analysis. In my experience, most learners show an extreme reluctance to produce any written texts at all.

Nor is it possible to simply work from already published experimental reports. Most of the published reports that use the LFP approach take for granted that the method works well, and it is very difficult to find an example of a published study that critically analyses the methodology. Again, this is not surprising. A widely acknowledged problem in experimental research is that studies which fail to show significant results are very hard to publish, and for this reason, studies where LFP appears not to be performing successfully are much less likely to be published than studies where it has been used successfully. In this way, the balance of published work will naturally tend to reflect the view that LFP is indeed a successful and effective method.

Fortunately, there is an alternative approach that we can use to evaluate some of the claims made about LFP. Rather than collect real data, we can model the way LFP works using simulated data, and we can use simulations of this sort to work out whether the models perform in the way they ought to do to if Laufer and Nation’s claims about LFP are correct. Studies of this sort are not common in applied linguistics, although they are a standard way of evaluating theories in psychology and engineering. This type of approach is often referred to as Monte Carlo analysis, because it typically relies on randomly generated data sets to model a complex process. (Early Monte Carlo studies sometimes used data in the form of random numbers produced by using a roulette wheel and, although it is now more usual to use data generated by a computer program, the name has stuck.) The advantage of using a Monte Carlo approach to the evaluation of LFP is that it gives us very great control over the characteristics of the texts we want to work with, and this allows us to explore the way LFP responds to these characteristics in a way that would be impossible with texts generated by real learners.

A series of studies which attempts to evaluate how LFP performs with a wide range of data sets is reported in the sections that follow.

2. SIMULATING LFP

The fundamental assumption underlying Laufer and Nation’s approach is that people with larger vocabularies generate texts which reflect this larger vocabulary, and this suggests that it ought to be possible to simulate the underlying experimental data by generating ‘texts’ from different sized vocabularies according to some rule. The obvious rule to use is simply to select words at random from a vocabulary. However, we know that words are not equally likely to occur in a text—a handful of highly frequent words account for a very large proportion of most texts, and the chances of rare words occurring in a text are vanishingly small. This suggests that we need to weight our selection of words in a way that reflects the effects of frequency. One simple way to do this is to use a model which applies a logarithmic
weighting to the selection of words. This has the effect of making highly frequent words much more likely to be selected, and rarer words much less likely to be selected for the ‘text’. The justification for using a logarithmic weighting, rather than any of the other possible weightings, comes from Zipf’s law, which states that there is a direct relationship between the frequency of a word in a large corpus and the log of its rank frequency (Zipf 1935; Herdan 1963). This sounds technical, but in practice it is much easier to understand what is going on than the description implies.

Table 2 shows a logarithmic transformation of the numbers 1,000 to 10,000. The details of this transformation are not important: what it is important to note is that the transformation stretches out the bottom end of the scale, and compresses the top end of the scale. Now let us suppose that we want to simulate the kinds of text that will be produced by a speaker with a vocabulary of 2,000 words. First, we calculate \( \text{ln}(2,000) \times 1,000 \). The answer to this calculation is 7,600. Next we generate a random number between 1 and 7,600. If the random number selected lies between 1 and 6,907, then a 1k word is added to our ‘text’. If our selected random number is between 6,908 and 7,600, then a 2k word is added to our ‘text’. If we repeat this procedure 300 times, then we can generate a 300 word ‘text’ with varying proportions of words in the differing frequency bands. Using this logarithmic transformation means that our text will contain a large proportion of 1k words, and a smaller number of 2k words: in this example, roughly 90 per cent of our text should be made up of 1k words. Of course, the text does not contain any real words—it is just a collection of numbers which has a distribution similar to what happens when words are produced in real texts.

The question we have to ask now is how realistic is this as a simulation? The top line in Table 3 shows the actual data reported in Laufer and Nation (1995). In their experiment, three groups of students were tested, each with differing levels of proficiency, and a mean profile for each group is reported. Broadly speaking, the profiles show that the percentage of 1k words in the profiles declines as proficiency increases, while the percentage of 2k words...
remains approximately constant. Laufer and Nation also make separate claims that the percentage of UWL words and NiL words in the profiles increases with increased proficiency. In Table 3, I have combined these two values into a single combined value which I have labelled 3k+. (The justification for this is that membership of the UWL list is not strictly determined by frequency.) The figures strongly suggest that the combined 3k+ scores rise as proficiency increases.

How do these real data compare with the data patterns produced by our simulations? The middle line of Table 3 shows data from three simulations produced by a recursive curve matching method. Basically, our model has one key parameter, Vocabulary size, and it generates a slightly different profile depending on the value of this parameter. By experimenting with different values of the Vocabulary size parameter, we can get a profile which matches the data in Laufer and Nation more or less closely. Curve matching tells us which setting for the Vocabulary size parameter gives the best fit with the data. In Table 3 we have three simulations, one for each of Laufer and Nation’s groups, in which I have tried to match data generated by the model with the actual data reported in Laufer and Nation’s paper. In the first simulation, we have Laufer and Nation’s low level group, and simulated data generated when the Vocabulary size parameter is set at 3,162. This generates a curve which closely mirrors the raw data.

In the second simulation, I have compared the raw data from Laufer and Nation’s second group with simulated data generated when the vocabulary size parameter is set at 6,230 words. Again, we have a close match between the real data and the simulation. In the third simulation, the Vocabulary size parameter is set at 8,255, and once again we get a profile which closely mirrors the real data generated by Laufer and Nation’s most advanced group.

These simulations parallel the data reported in Laufer and Nation (1995) to an astonishingly close degree. For all practical purposes, the simulations and the real data are indistinguishable, and this suggests that we might be able to use the best fit vocabulary size parameter as an estimate of the vocabulary size of Laufer and Nation’s subjects. Laufer and Nation describe their first set of subjects as ‘low intermediate level’. It is not exactly clear what this description means, but later information implies that these subjects have not yet taken the Cambridge FCE examination. As a broad rule of thumb, my research group reckons that students at this level typically have a

### Table 3: Data from Laufer and Nation (1995) and a set of best-fit simulations

<table>
<thead>
<tr>
<th></th>
<th>1k</th>
<th>2k</th>
<th>3k+</th>
<th>1k</th>
<th>2k</th>
<th>3k+</th>
<th>1k</th>
<th>2k</th>
<th>3k+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laufer and Nation’s data</td>
<td>86.5</td>
<td>7.1</td>
<td>6.4</td>
<td>78.0</td>
<td>6.7</td>
<td>13.7</td>
<td>77.0</td>
<td>7.0</td>
<td>16.0</td>
</tr>
<tr>
<td>Simulation</td>
<td>85.7</td>
<td>8.6</td>
<td>5.7</td>
<td>79.7</td>
<td>6.7</td>
<td>13.6</td>
<td>76.6</td>
<td>7.7</td>
<td>15.7</td>
</tr>
<tr>
<td>Vocabulary size estimate</td>
<td>3,162</td>
<td>6,230</td>
<td>8,255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
receptive vocabulary of around 4,000, but their productive vocabulary is likely to be smaller than this, and 3,162 seems to be a reasonable estimate for the productive vocabulary of a group of subjects at this level.

Laufer and Nation’s second and third groups had both passed an FCE examination, and taken either one or two further semesters of English course work. Here, the best fitting simulations suggest that these groups have a vocabulary of 6,230 words and 8,255 words respectively. These figures strike me as somewhat on the high side, since they imply that subjects can improve their vocabulary by 3,000 or more words in the space of only one semester. However, the subjects in the three groups are actually not directly comparable, since the first group is a mixed L1 group following courses in New Zealand, while the other two groups are linguistically homogeneous, and are based in Israel. Clearly, this makes it difficult to extrapolate from the first group to the other two. For the moment, however, we will take these vocabulary estimates at face value, and guess that Laufer and Nation’s first group has a vocabulary size of about 3,000 words, while the second group has a vocabulary size of around 6,000 words, and the third group has a vocabulary size of about 8,000 words. Clearly the simulation model generates data which are not wildly dissimilar from the real data generated by learners, and this supports our view that we can use the simulations to test whether LFP really works.

Laufer and Nation’s groups consisted of just over 20 subjects. They asked these subjects to produce two essays, each of about 300 words. They then derived Lexical Frequency Profiles from each of these texts and used these data to ask a number of questions designed to evaluate the performance of LFP as a research tool.

Their first question is: ‘Will there be a significant difference between the LFPs of learners of different language proficiency levels?’ This comparison is fairly straightforward, and involves simply comparing the mean scores of the groups in an ANOVA. Laufer and Nation report that there is indeed a significant group effect in the data. They actually report a series of F-tests, one for each of the frequency levels, and report that significant differences exist for the proportion of 1k words, the UWL words and the NiL words, but not for the 2k words.

Further analysis suggests that this difference is largely due to Group 1’s performance being worse than that of the more advanced groups, which did not differ from each other. In spite of this, Laufer and Nation later assert, without any proviso, that LFP ‘discriminates between learners of different proficiency levels’.

How can we evaluate this claim using the simulation method? The simplest approach is to generate a set of 300-word ‘texts’ using the sampling procedure described above. We can approximate Laufer and Nation’s
Group 1 by running a set of 20 simulations where the vocabulary size parameter of each simulation is about 3,000—in practice, we need to allow this figure to vary by $+/\text{-}10$ per cent in order to simulate the kind of range that you would get in a real group of subjects. We can then compare these data with a simulated data set generated on the assumption that we are dealing with a group of subjects with a vocabulary size of 3,500 words. If we repeat this exercise 100 times, and count the number of times we come up with a ‘significant’ result, then the simulations will give us a rough idea of how reliable Laufer and Nation’s claim is. Clearly, if their claim is correct, then close to 100 per cent of the simulations should produce a ‘significant’ result. Having done this computation, we can then go on to compare data generated from a 3,000-word vocabulary with data generated from a 4,000-word vocabulary, a 4,500-word vocabulary, a 5,000-word vocabulary, and so on. Repeating this procedure over a wide range of values will give us a good idea of how large the difference between two vocabularies needs to be in order for a reliable difference in the Vocabulary Profiles to emerge. Data from a set of simulations of this sort is shown in Table 4.

Table 4 summarizes a very large set of simulated experiments. For each experiment I generated LFP profiles based on a set of forty 300-word texts: two sets of twenty texts, with the Vocabulary size parameter set at a different value for each text set. I then compared the profile sets and looked for a significant difference between the two sets. Each cell in the table represents 100 simulations of this sort. The first cell in line two summarizes 100 simulations in which I compared profiles generated by a 3,000-word vocabulary with profiles generated by a 2,000-word vocabulary. A series of t-tests showed that the profiles generated by this procedure were always significantly different from each other. The second cell summarizes a further 100 simulations which compared the profiles generated by a 3,000-word vocabulary and a 2,500-word vocabulary. In this case, only 90 per cent of the simulations produce a statistical difference. The fourth cell summarizes 100 simulations comparing the profiles generated by a 3,000-word vocabulary and a 3,500-word vocabulary. Here the data suggest that only 67 per cent of these comparisons produce a significant difference. And so on for the other cells. In total, I ran a series of 5,100 simulations in this way—it would clearly be impossible to run experiments on this scale with real

| Words | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 7.0 | 7.5 | 8.0 | 8.5 | 9.0 | 9.5 | 10.0 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| V=3.0k | 100 | 90  | 67  | 98  | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| V=6.0k | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 95  | 67  | 23  | 25  | 42  | 75  | 88  |
| V=8.0k | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 99  | 87  | 70  | 36  | 12  | 20  | 28  | 54  | 66  |
human subjects. The table shows that, generally speaking, LFP profiles are reliably different as long as the vocabularies that they are generated from differ substantially in size. It is very hard to distinguish between the profiles generated by two vocabularies which are similar in size.

A number of assumptions had to be made to generate the data in Table 4. First, I have set a very lenient criterion ($p = .05$) for deciding whether a result was ‘significant’ or not. This decision is an arbitrary choice, which maximizes the number of ‘significant’ results reported. Choosing a stricter criterion ($p < .01$ or $p < .001$) would have reduced the number of significant results. Secondly, the computations are based on the percentage of Level 1 words contained in each ‘text’. This is a slight over-simplification of Laufer and Nation’s data, but not a large one. Laufer and Nation actually compute difference scores for the 1k words, the 2k words and the higher bands separately. Their data show that the groups do not differ in the 2k band, however, and since we have collapsed the higher bands into a single figure, there is an inverse relation between the 1k scores and the 3k+ scores generated by the simulations. For all intents and purposes, then, the 1k values are a reasonable and convenient shorthand for the entire data set.

The simulated data are actually very close to Laufer and Nation’s data in those cases where direct comparisons can be made. The simulations suggest that it will always be possible to distinguish between a text generated by a 3,000-word vocabulary and one of 6,000 or 8,000 words, and this result is in line with Laufer and Nation’s report. At higher levels, the data are more difficult to interpret: the table suggests that it will not always be possible to distinguish between texts produced by a 6,000-word vocabulary and an 8,000-word vocabulary. Generally speaking, there will be a difference between these texts, but a small proportion—about 13 per cent of trials—will fail to show a difference. Again, this is broadly in line with Laufer and Nation’s data. However, what is striking about the figures is that reliable differences emerge only when the two vocabularies differ by a substantial amount. The model suggests that it is not possible to distinguish between two vocabularies differing by only 500 words: for smaller vocabularies there is more tolerance, but for larger vocabularies it is increasingly difficult to find reliably different results. Profiles generated by an 8,000-word vocabulary are reliably different from profiles generated by a 5,000-word vocabulary, but not reliably different from the profiles generated by a 6,000- or a 7,000-word vocabulary.

Laufer and Nation’s conclusion that ‘LFP discriminates between learners of different proficiency levels’ is partly true, then, but the simulations suggest that we need to add some very strong riders to this claim. LFP looks as though it might discriminate reasonably well between lower level learners and very advanced learners, but it is much less good at distinguishing between groups that are more evenly matched. Certainly, if we adopt Milton and Meara’s (1998) rule of thumb that L2 vocabularies typically grow by about 500 words for every year of study, it seems unlikely that LFP would be
able to discriminate reliably between two adjacent year groups. In addition, any relaxation of the parameters—adopting a stricter criterion for ‘significance’, or allowing the simulated groups to be less homogeneous—would enlarge the zones where we fail to find 100 per cent significant results. This picture is rather more subtle than the one painted by Laufer and Nation’s account of LFP.

4

Laufer and Nation’s second question is ‘Will the LFP of the compositions correlate highly with the scores of the same learners on the active version of the Vocabulary Levels test?’ This is basically a claim about a direct relationship between LFP and the underlying vocabulary size of the Ss, and what Laufer and Nation are asking here is whether a measure of active vocabulary will correlate with the profiles. In a further paper, Laufer and Nation (1999) developed a test—the Active version of the Vocabulary Levels Test—which they claimed could be used to estimate productive vocabulary size. The details of this test do not concern us here; however, the underlying logic of the argument does. For obvious reasons, Laufer and Nation do not have any direct measure of the vocabulary sizes of their subjects, so they need to use something like the Vocabulary Levels Test as a way of estimating this figure. Tests of this sort always introduce a degree of uncertainty into the data, because they can only sample the words people know and extrapolate from these samples. This means that Laufer and Nation’s vocabulary size measure is at best a rough approximation to the real vocabulary size of their subjects. In our simulations, however, we know exactly how big the vocabularies we are dealing with are—this figure is determined by our vocabulary size parameter—so we can compare LFP scores directly with vocabulary size.

We can simulate Laufer and Nation’s data in the following way.

- First, we generate a group of subjects (S) with a mean Vocabulary Size of 2,000 words, allowing this figure to vary by a small percentage up or down in order to reflect natural variance in a group of learners. Then for each S we record the simulated vocabulary size. In this way, if we allow our Vocabulary size scores to vary by 10 per cent, we get a group of subjects whose Vocabulary size scores vary between 1,800 and 2,200.
- Next we generate an LFP score for each S, using the procedure outlined above.
- Finally, we correlate the vocabulary size scores with the LFP profile scores.
- We repeat this procedure 100 times, and count the number of ‘significant’ correlations the simulations produce.
- We then repeat all these steps for a group of subjects with vocabularies of 2,500, 3,000, 3,500, 4,000 words and so on.

Laufer and Nation report that their data show a good correlation between their LFP scores and the scores of the subjects on the active version of the
Levels Test ($r = .7$, $p < .0001$). The size of this effect suggests that it must have been achieved by combining all the subjects into a single group, effectively creating a group of 60 subjects, with a mean vocabulary size of around 5,000 words and a correspondingly large standard deviation. We can simulate this set up by running a series of simulations where $N = 60$, Mean Vocabulary Size $= 5,000$, and the individual subjects are allowed to vary by 50 per cent from this average figure. Approximately 95 per cent of the simulations we ran with this combination of parameters do indeed produce correlations in the region of .7, which again confirms that our simulation model is a reasonable approximation of Laufer and Nation’s model, in spite of the simplifications we have built into it. However, it is notable that Laufer and Nation do not report the correlations for the smaller groups, and it seems very unlikely that smaller, more homogeneous groups would produce such clear cut results.

In Table 5, I show the number of ‘significant’ correlations produced when we run the simulations 100 times using a range of plausible assumptions about the amount of variation in a group. As in Table 4, each cell summarizes 100 simulations. In each of these simulations I have generated a set of 20 profiles. For these profiles, the vocabulary size parameter is allowed to vary within certain limits, in order to reflect the fact that a group of real human subjects always shows some variation. In cell one row two of Table 5, for example, we have results from a set of profiles based on a 2,000-word vocabulary, but this figure is allowed to vary up or down by up to 10 per cent. In the cell below that, we have results from a set of profiles based on a 2,000-word vocabulary, but this figure is allowed to vary up or down by 25 per cent. In this way, the second row of Table 5 simulates groups of homogeneous subjects, while rows three and four of Table 5 simulate more heterogeneous subject groups. In this table, ‘significant’ means a correlation where $p < .01$. The figures are based on a group size of 20, with the vocabulary size of the groups ranging from 2k to 10k, and with three levels of variation within the group. The data show that there is a strong tendency for significant results to become less likely as vocabulary size increases. Furthermore, the chances of a significant result are strongly affected by the variation in vocabulary size across the group. If vocabulary size within the group varies by $\pm 25$ per cent then there is a modestly good chance of

| Vsize (k) | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 6.5 | 7.0 | 7.5 | 8.0 | 8.5 | 9.0 | 9.5 | 10.0 |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| +/-10%    | 29  | 19  | 16  | 10  | 7   | 10  | 5   | 10  | 5   | 3   | 7   | 4   | 7   | 2   | 7   | 6   |
| +/-25%    | 92  | 91  | 76  | 65  | 67  | 53  | 48  | 50  | 47  | 38  | 37  | 47  | 41  | 33  | 24  | 28  | 34  |
| +/-50%    | 100 | 100 | 100 | 100 | 97  | 99  | 99  | 98  | 95  | 97  | 95  | 97  | 92  | 97  | 95  | 95  | 89  |
returning a significant correlation between Vocabulary size and the LFP score, as long as the mean vocabulary size of the group is small, although even under these conditions, the number of significant results falls well short of 100 per cent. Where the test groups are more homogeneous, then it is extremely difficult to obtain a significant result.

This suggests that Laufer and Nation’s claim that LFP ‘correlates well with an independent measure of vocabulary knowledge’ overstates the facts somewhat, and that it would be unwise to assume their results would generalize other than to very large groups comprising students with a wide range of vocabulary sizes.

5

Laufer and Nation’s third question is ‘Will the LFPs in two sets of compositions written by the same learners correlate highly with each other?’ This question is one which can easily be simulated by the model, since it is essentially the same problem as the one we discussed in the previous section: the only change we need to make is to replace the Vsize data with a second LFP score. In these simulations, then, we carry out the following steps.

- First we generate a group of subjects with a mean vocabulary size of 2000 words, allowing this figure to vary by a small percentage for each subject.
- Next we generate two LFP scores for each subject and correlate the two sets of scores.
- We repeat this process 100 times.
- We then repeat this entire procedure for other values of Vsize and for other values of the homogeneity parameter.

Laufer and Nation’s model implies that these correlations should be high, and that the number of significant correlations in a run of 100 simulations should be close to 100 per cent. Table 6 shows the results of a set of simulations which test this claim.

As in Table 5, each cell summarizes the results of 100 simulated experiments. In each experiment, we compare 20 pairs of profiles. The score in the cell shows what proportion of these simulations produce a significant correlation.

Table 6: Percentage of ‘significant’ correlations (p < .01) between two LFP profiles for various parameter combinations

| Vsize (k) | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 | 5.0 | 5.5 | 6.0 | 6.5 | 7.0 | 7.5 | 8.0 | 8.5 | 9.0 | 9.5 | 10.0 |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| +/-10%   | 6   | 9   | 1   | 1   | 0   | 1   | 4   | 0   | 0   | 3   | 1   | 3   | 2   | 1   | 0   | 1   |
| +/-25%   | 57  | 39  | 28  | 20  | 15  | 8   | 5   | 8   | 6   | 6   | 4   | 7   | 4   | 4   | 4   | 5   | 8   |
| +/-50%   | 99  | 99  | 97  | 92  | 92  | 90  | 87  | 92  | 83  | 67  | 75  | 71  | 69  | 60  | 54  | 61  | 58  | 49  | 43  |
correlation between the profile sets. Thus, cell 1 in the first row shows that only 6 per cent of the simulations produce significant correlations.

Since these simulations are essentially the same as the simulations reported in Table 5, but with an added degree of uncertainty in the data, it is not surprising that these data show broadly similar patterns, but rather worse outcomes. As in Table 5, significant results are more likely to be found where the mean vocabulary size of the group is small, and when the variation within the group is large. However, significant results occur reliably only when there is an extreme spread of scores within the group, and the group’s mean vocabulary score is low.

Interestingly, Laufer and Nation do not actually report the correlations that their data produce, which suggests that they too failed to find the significant correlations that they were expecting. Instead, they fall back on an analysis in which they compare the mean LFP scores produced by the group for two essays using t-tests. This analysis failed to generate a significant difference for the two essays, they argue, and thus they conclude that LFP is broadly stable. This is clearly a very weak claim: it is basically a null hypothesis, with a very high probability of being confirmed by chance data. As such, it somewhat undermines Laufer and Nation’s claim that LFP provides similar stable results across two pieces of writing by the same subject.

Laufer and Nation’s fourth question is: ‘Will the percentages of words at each frequency level correlate highly with each other in the two sets of compositions?’ I think this question is a refinement of question 3, but focused on the separate levels rather than on the overall profile, but it is difficult to be certain of this as Laufer and Nation do not in fact address this question directly in their analysis. Instead, they develop a new measure which discards any 1k words occurring in the texts, and generates a second profile based on the remaining words. Thus, a text which consisted of 300 words, of which 250 were level 1k words would have a remainder of 50 words. If these were proportioned as 25 2k words, 15 UWL words and 10 NiL words, then the new profile would be 50 per cent, 30 per cent, 20 per cent. Laufer and Nation claim that a set of t-tests shows no significant differences between the essays, and they use this to reinforce their view that the LFPs are stable. They actually argue that this revised ‘profile is a better measure of lexical richness in general’, particularly for more advanced students.

The problem with this analysis is that again we have a null hypothesis masquerading as a real one. If LFP is stable, then two essays produced by the same subject would be expected to produce the same profile, and it is simply inappropriate to evaluate this claim by looking for non-significant t-tests. A better way would be to argue that very different sized vocabularies ought
to produce large differences in their profiles, while similar sized vocabularies will not. We can test this claim using the simulator. We do this as follows.

- First we generate data for two groups of 20 subjects, one with a vocabulary of 2,500 words, the other with a vocabulary of 3,000 words.
- Next we compute the total of words not in the 1k list.
- Then we compute a profile that measures the proportion of 3k+ words in this shortened text.
- We then compute whether there is a significant difference between the two groups.
- We then repeat this procedure 100 times.
- After that, we generate data from two groups, one with a vocabulary of 2,500 words, the other with a vocabulary of 3,500, and repeat the procedure.
- And so on, increasing the vocabulary size of the groups at each iteration of the process.

Laufer and Nation’s analysis implies that we will get very small, non-significant differences between the groups when they have similar vocabulary sizes, and much larger and significant differences between the groups when the vocabulary size differences are large. The relevant data for testing this claim are shown in Table 7.

The table clearly shows that the proportion measure does NOT tend to generate significant differences, even when the source vocabularies are hugely different. It is therefore not at all surprising that Laufer and Nation failed to find a significant difference between texts generated by the same subject, and by the same token, their claim that the proportion measure is stable across texts produced by the same author must be treated as a somewhat generous interpretation of their actual data.

### Table 7: Percentage of ‘significant’ results (p < .01) in a run of 100 simulations comparing different size vocabularies using the revised profile method

<table>
<thead>
<tr>
<th>Vsize (k)</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
<th>4.5</th>
<th>5.0</th>
<th>5.5</th>
<th>6.0</th>
<th>6.5</th>
<th>7.0</th>
<th>7.5</th>
<th>8.0</th>
<th>8.5</th>
<th>9.0</th>
<th>9.5</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>3000</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

7. CONCLUSIONS

This paper has used a set of Monte Carlo simulations to evaluate a number of strong claims made by Laufer and Nation about the Lexical Frequency Profile procedure. Each simulation consists of several thousand pseudo-experiments,
using different combinations of a few simple parameters, and this amounts to a much more robust evaluation of the LFP procedure than could possibly be developed from a real experiment. The simulations suggest that Laufer and Nation’s claims may be a lot less robust than they are made out to be. They strongly suggest that LFP does not reliably distinguish between learner groups at different levels of proficiency, probably does not produce strong correlations between scores on different texts produced by the same author, and that the reported consistency of LFP profiles from one administration to another may be an experimental artefact.

This, of course, is a harsh evaluation of the LFP procedure, on purely technical grounds. The simulations reported in the paper all depend on the assumption that it is indeed plausible to model the production of texts using the logarithmic model described in Section 2. I have argued here that, despite the obvious simplifications of the model, it is a plausible way of simulating the data we need to evaluate the LFP, but this is clearly an empirical question that needs further study.

The criticisms of LFP that have been developed here should not, of course, be taken as a criticism of the underlying questions that motivated Laufer and Nation’s 1995 paper. In my view, Laufer and Nation are absolutely right to identify the assessment of productive vocabulary as a crucial element in developing theories of vocabulary development in L2 learners. As long as we do not have reliable tools which allow us to make meaningful statements about the relationship between active and passive vocabulary in L2 learners, it will not be possible for us to test any of the models which make claims about how this relationship develops, and what effects it might have on other aspects of L2 performance. Claims of this sort are becoming increasingly common in the literature, and it is very important that we develop really good tools that will enable us to evaluate these claims properly.

I hope that this paper will stimulate further debate about the kind of tools we need to measure productive vocabulary meaningfully, and lead to further development of the very important ideas which Laufer and Nation’s paper begins to address.

Final version received June 2004

REFERENCES


