_lognostics
tools for vocabulary research

P_Lex: a simple and effective way of describing the lexical characteristics of short L2 texts.

Paul Meara ***Swansea University***
Huw Bell ***Manchester Metropolitan University***

Introduction

There are a number of reasons why it would be useful to have a formal measure of the lexical characteristics of L2 texts. The most obvious of these is that teachers are frequently called upon to make judgements about the types of vocabulary their students use, and to judge the adequacy of this vocabulary. Recent work, for example, by Engber (1995) suggests that these judgements are actually an important component in the way teachers make overall assessments of texts that L2 learners produce. For example, Engber found that a significant proportion of the oral marks awarded to L2 texts could be accounted for by a vocabulary rating. Assessments of this sort also play an important part in the public examinations system, with many examinations requiring examiners to assess explicitly the vocabulary used by candidates. To a very large extent, however, these assessments rely on subjective judgements. Clearly, it will be helpful if we could be more specific about exactly what constitutes 'wide' or 'adequate' vocabulary.

The most common approach to this problem has been to work with measures of lexical richness, or lexical diversity. Indices of this sort have been widely used in studies of lexico-statistics (e.g. Herdan 1960) and L1 development (e.g. Miller and Klee 1955), and they have begun to appear with increasing frequency in work on L2 speakers (e.g. Arnaud 1984, Broeder et aL. 1988, Malvern and Richards 1977). Most of this work uses measures that compare the number of Lexical Types in a text with the number of Lexical Tokens in the same text. A number measures of this sort exist (see table 1), but there is no clear agreement about which is the best variant to use in the context of L2 learners. Indeed, there is some considerable disagreement about exactly what features of texts these methodologies describe. The main practical difficulty with measures based on Types and Tokens is that they are sensitive to the length of the text being assessed (longer texts typically have lower Type Token Ratios than shorter ones do), and this makes things difficult in real assessment situations, where it is often hard to control the length of the texts you need to assess. Malvern and Richards (1977) have suggested that their measure D is largely insensitive to length, but this measure has not been widely taken up in L2 studies yet, partly because it is computationally more complex than the other measures listed in table 1.

The measures listed in table 1 are all examples of what we might call *intrinsic measures of lexical variety*. In these measures, variety is assessed solely in terms of the words that appear in the

Table 1: The principal measures of lexical variation based on Types and Tokens

Measure	Name of Measure
N	number of Tokens in a Text
V	number of Types in a Text
V1	number of 'hapax legomena' – words occurring only once in a text
V/N	Type/Token Ratio
V/√N	Guiraud's Index
logV/logN	Herdan's Index
V^t	theoretical vocabulary

text itself. The critical variables here are the number of Tokens and the number of Types in the text, and no attempt is made to categorise these items according to criteria that are external to the text itself. This strikes us as an odd way to go about evaluating lexical resources of L2 speakers. Put in the simplest terms, any analysis that relies on Types and Tokens will produce identical scores for the examples below:

example 1: the man saw the woman.
example 2: the Bishop observed the actress.
example 3: the magistrate sentenced the burglar.

Each example consists of five words, two of which are repeated, giving a total of five Tokens and four Types. Whichever variant of the Type/Token ratio we adopt, these data are identical from the point of view of a Type/Token analysis. Intuitively, however, the three examples are quite different. Example 1 uses vocabulary that is very common, while examples 2 and 3 use more unusual vocabulary. It would be difficult to say anything very significant about the vocabulary resources of a learner who produced sentence 1, but sentences 2 and 3 are unlikely to have been produced by beginners with limited vocabulary resources.

Notice, however that we are now making a judgement that is not based simply on the evidence available in the text itself. We consider the vocabulary in example 1 to be simple, because we know about the frequency characteristics of words in the language as a whole. This knowledge allows us to say that *man* and *woman* are 'easy words' -- words that are frequent in English, and that we would expect most speakers in English to know -- whereas *bishop* and *actress* or *magistrate* and *burglar* are not, and we can use this information to make some fairly strong inferences about the total lexical resources that are available to the writer.

This suggests that there might be a case for developing some *extrinsic measures of lexical richness* for use with L2 learners. These measures would not be limited to the number of Types or Tokens appearing in an L2 text: they would supplement this information with additional information about the sort of words being used, and the sorts of lexical choices that are

being made in particular text.

An example of a measure of this sort is to be found in Laufer and Nation's *Lexical Frequency Profile* (Laufer and 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 and Nation 1984); category 4 includes any word not found in the previous three lists. An example of the output from LFP can be seen in table 2, which contains an LFP analysis of this paragraph.

Table 2: Part of a Lexical Frequency Profile analysis

Word list	**Tokens/%**	**Types/%**	**Families**
One	100/76.3	51/73.9	48
Two	10/ 7.6	6/ 8.7	4
Three	15/11.5	8/ 11.6	7
Not in Lists	6/ 4.6	4/ 5.8	???
Total	**131**	**69**	**59**

Number of BASEWORD1.DAT types 2804 Number of BASEWORD1.DAT families: 958
Number of BASEWORD2.DAT types 2614 Number of BASEWORD2.DAT families: 1028
Number of BASEWORD3.DAT types 2847 Number of BASEWORD3.DAT families: 836

Note: this table shows the data in the format generated by the LFP programme.
The figures should be interpreted as follows:
line 1 indicates that 100 word tokens can be found in Nation's level one word list. This figure represents 76.3% of the total word token count. Fifty-one word types can be found in Nation's level one list. This figure represents 73.9% of the total type count. When these words are collapsed into their appropriate word families -- e.g. by treating *happy, unhappy, happiness* as a single word family -- then we are left with 48 distinct word families from Nation's level one list.

Lines 2 and 3 are to be interpreted in a similar way.

Line 4 indicates that LFP was not able to process six word tokens (4.6% of the total) because the words are not in Nation's list. (This usually means that the words in question are very low-frequency words.) In terms of word types, these four items made up 5.8% of the total type count. LFP is not able to assign these words to word families, since it did not recognise them.

The last three lines of the output report how many words LFP recognises at each of its three levels. Level 1, for example, contains 2804 word types, which LFP classifies as belonging to 958 different word families.

Laufer and Nation make a number of strong claims for LFP. Specifically, they claim that the ratio of category 3 and category 4 words to category 1 and category 2 words provide a reliable index of the vocabulary resources available to an L2 writer. They claim that these LFP scores are stable over time, correlate closely with proficiency measures, and are

relatively unaffected by task. That is, Laufer and Nation claim that L2 writers have characteristic LFP scores, which are reliably stable over a number of different test conditions:

> The LFP has been shown to be a reliable and valid measure of lexical use in writing. It provided similar stable results for two pieces of writing by the same person, and discriminate between learners at different proficiency levels. It correlates well with an independent measure of vocabulary knowledge. (Laufer and Nation 1995: 319)

It is not our intention here to quarrel with Laufer and Nation's interpretation of their results -- though it might be worth pointing out that our own experience with LFP-type measures suggests that they are much less reliable and much less sensitive than Laufer and Nation claim. In our experience, LFP has poor measurement characteristics, and does not discriminate well between texts, because it relies very heavily on a simple count of the category 3 and category 4 words in the text. The number of these words in a 'typical' text is usually very small, and this severely limits the way LFP works. In practice, the percentage of category 3 and category 4 words in a text rarely exceed 10 percent, so the range of scores produced by LFP is fairly limited, and we think that this lack of variation may be the biggest contributor to Laufer and Nation's stable scores. A particular problem arises with text produced by very low level learners, when the percentage of category 3 and category 4 words is often close to zero.

More importantly, a serious practical problem with LFP is that it requires relatively long texts for stable measures to emerge. Laufer and Nation claim that in their data 'profiles over 200 words were found to be stable, while those done on less than 200 words were not' (1995:314). This seems to us to be a very serious practical limitation. 200 words is a substantial amount of text, and our experience suggests that it is very difficult to extract texts of this length from learners unless they are relatively advanced, or unusually cooperative. LFP appears, in any case, to be very sensitive to text length, and this makes it difficult to compare LFP scores from different sources.

It seems, then, that there might be a case for developing an alternative measure that, like LFP, uses extrinsic characteristics of words to evaluate the vocabulary resources of their authors, but has the additional advantage that it works with short texts and produces scores with good measurement characteristics. A measure of this sort is described in the next section.

P_Lex

P_Lex is based on the idea that it might be possible to make a virtue out of the fact that 'difficult' words occur only infrequently in texts. P_Lex looks at the distribution of difficult words in a text, and returns a simple index that tells us how likely the occurrence of these words is. The underlying assumption here is that people with big vocabularies are more likely to use in frequent words than people with smaller vocabularies are, and that we can use

the index we derive from the texts as a pointer to vocabulary size. P_Lex is a first step in this direction.

P_Lex works as follows. Suppose we want to process text T. First, we divide T into a set of 10 word segments, ignoring punctuation. Next, we categorise the words in each segment in terms of their objective frequency. The current version of P_Lex is based on Nation's 1984 word lists. It treats all words occurring in the first 1000 word list as 'easy'. Proper nouns, numbers and geographical derivatives are also categorised as 'easy' words. All other words are categorised as 'hard'. Next, we count the number of infrequent words in each segment, and calculate the number of segments containing zero infrequent words, the number of segments containing one infrequent word, the number of segments containing two infrequent words, and so on. This gives as a P_Lex profile for text T. We can make this a bit more concrete with a real example. Consider the following text:

> I come from Japan. My home town is Okinawa. Is in the south of Japan, and there is a very big American air-base. My father is engineer at home. I come to Swansea to study engineering, like my father did. But he is teaching engineer and I want to be real scientist. My teacher wants me to work on a new kind of protein found only in seaweed. You have lots of this seaweed in Swansea, so this is a good place for me to come. My journey was very long, and I am very tired now. I have been to Wales before when I was a boy.

The text contains a total of 108 words, which P_Lex splits into 10 ten-word segments like this:

1: I come from Japan. My home town is Okinawa. Is	0
2: in the south of Japan, and there is a very	0
3: big American air-base. My father is engineer at home. I	2
4: come to Swansea to study engineering, like my father did.	1
5: But he is teaching engineer and I want to be real	1
6: scientist. My teacher wants me to work on a	1
7: new kind of protein found only in seaweed. You have	2
8: lots of this seaweed in Swansea, so this is a	1
9: good place for me to come. My journey was very	0
10: long, and I am very tired now. I have been	0

The first and second segments contain no hard words. The segment three contains two hard words (*air-base* and *engineer*). Segments 4,5 and 6 containing one hard word each (*engineering*, *engineer*, *scientist*). Segments 7 contains two hard words (*protein*, *seaweed*). Segment 8 contains one hard word (*seaweed*). Segments 9 and 10 contain no hard words. We use these data to construct a P_Lex profile, like the one shown in table 3. Table 3 simply tells us that four of the segments (40%) contain no hard words, four segments (40%) contain one hard word,

and two segments (20%) contain two hard words. No segments contain three or more hard words.

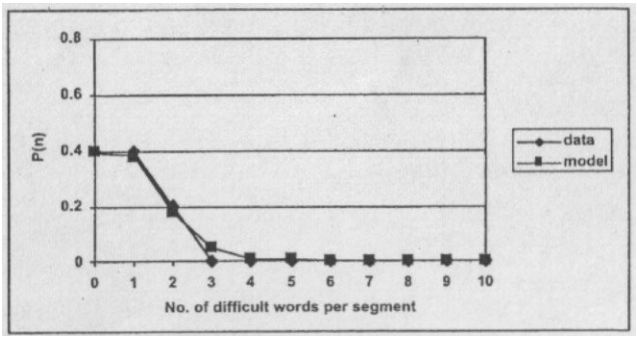
Table 3: number of segments in the text that contain N difficult words.

N	0	1	2	3	4	5	6	7	8	9	10
nmbr	4	4	2	0	0	0	0	0	0	0	0
prop	.4	.4	.2	0	0	0	0	0	0	0	0

Nmbr = the raw number of segments containing N hard words
prop= the same figures interpreted as proportions

Not surprisingly, it turns out that the distributions we get for this type of analysis are strongly skewed to the left: most texts contain few difficult words, and texts that contain a very high proportion of such words are themselves quite unusual. Distributions that are strongly skewed to the left are often well described by so-called *Poisson distributions*, and this turns out to be the case with the data that our analysis produces. Figure 1, for example shows the data from table 3 together with the Poisson curve that fits this data very closely.

Figure 1: Matching data to a theoretical curve, lambda =0.92.



The mathematics of fitting curves to data is fairly complex, so we have summarised this process in detail in Appendix A. For readers who don't want to get that involved, it is enough to know that there is a procedure which makes it possible to turn data like that in Table 3 into a single figure, conventionally known as *lambda*. Our contention is that these lambda values provide much tidier data than the ratio figures that are produced by LFP. The lambda figures are much easier to work with than the LFP ratios, and are much more intuitive to interpret. The lambda values typically range from 0 to about 4.5, with higher figures corresponding to a higher proportion of infrequent words. Lambda values have good measurement characteristics, and this allows them to be added and averaged straightforwardly. More importantly, however, lambda scores are much less sensitive to text

length than the LFP scores are, and, critically, the P_Lex methodology gives lambda scores that are reasonably stable with very short texts.

An evaluation of P_Lex

This section illustrates the way P_Lex works with a large set of texts produced by L2 learners in English. It addresses the issue of how reliable P_Lex scores are and how well the scores correlate with other measures of productive vocabulary in L2. We also consider how well P_Lex works with texts and different lengths.

METHOD

A total of 49 subjects took part in this study. All were learners of English as a foreign language, taking part in summer courses at the University of Wales Swansea. These learners came from a variety of L1 backgrounds, and they exhibited a range of proficiency, ranging from lower intermediate to advanced.

Each subject produced two pieces of written work. For the purposes of comparison, we asked subject to produce two discursive essays, using the same titles as were used in Laufer and Nation's study (1995: 320) -- *Should a government be allowed to limit the number of children and family can have?* and *A person cannot be poor and happy: Discuss*. We also set the same conditions as Laufer and Nation did: each essay was to be written in an hour, without dictionaries. We asked subjects to produce about 300 words, though in practice many subjects produced essays that were shorter than this. The two essays were written within a week of each other, and we assume that this time lapse was too short for any real linguistic gains to have taken place in the subjects' vocabulary knowledge. Each student also took the active version of the Vocabulary Levels Text (Laufer and Nation 1999).

The essays were transcribed into machine readable format. Minor spelling errors were corrected at this stage. Following this, the essays were processed using the P_Lex methodology described above. The shortest essay included in the analysis consisted of just over 250 words, and we therefore used the first 250 words of each essay in the analysis that follows.

ANALYSIS

Our basic question was whether the P_Lex methodology is reliably stable across administrations, and in order to evaluate this question we calculated the lambda values for each subject's two essays. If the lambda values are reliable, then there should be no difference between the mean lambda values of the two essays, and across the group as a whole there should be a close correlation between the two sets of lambda values. The analysis confirmed that there was not a significant difference between the mean lambda scores, and that the two sets of scores correlated modestly. See table 4.

This is a reasonably good result. The data confirm that the P_Lex scores on essay 1 and essay 2 do not differ significantly, and there is a modest correlation between the P_Lex

Table 4: mean lambda scores and correlations – 250 word texts.

Essay 1:	***mean*** 1.466	**Essay 2:**	***mean*** 1.309
	sd. 0.56		***sd.*** 0.51

correlation – essay 1 and essay 2: r=0.655 p<.001

scores on the two essays. The correlations account for about 43 percent and the total variance (r=0.655). These figures are all broadly in line with the results reported by Laufer and Nation (1995).

We also examined whether the P_Lex measure was able to distinguish reliably between groups of learners at different levels of proficiency. We used the results of the Levels Test to divide the subject base into two subgroups: Group H comprised the top 24 subjects on this measure, and Group L comprised the bottom 25 subjects. Independent t-tests indicated that the P_Lex scores of these two groups were reliably different (t=4.69, p<.01 for essay 1; t=2.79, p<.01 for essay 2).

Table 5: Mean P_Lex scores for group H (high scorers) and group L (low scorers).

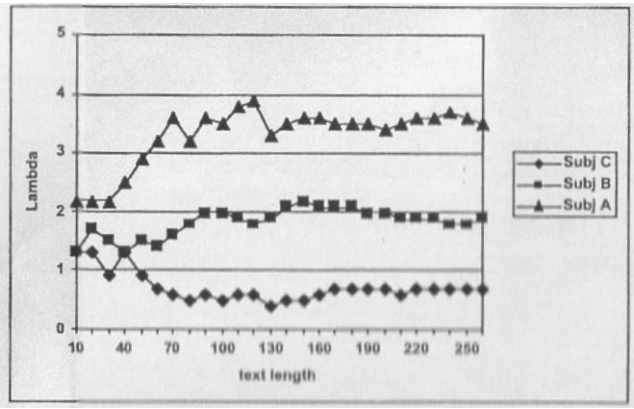
	Essay 1 250 wds		**Essay 2 250 wds**	
	Gp H	**Gp L**	**Gp H**	**Gp L**
Mean	1.785	1.160	1.503	1.124
sd.	*0.583*	*0.429*	*0.500*	*0.451*

The overall correlation between P_Lex scores and the Levels Test were also good. Essay 1 and the Levels Test correlated moderately and significantly (r=.565, p<.001), and essay 2 and the Levels Test correlated more modestly, but still significantly (r=.339, p=.017). A similar pattern of correlations between two pieces of written work and the Levels Test results was also found in Laufer and Nation's 1995 report.

These findings, then, broadly replicate the pattern of findings reported in Laufer and Nation (1995). A further analysis of the data showed that this was not a fluke: an analysis of all the texts using Laufer and Nation's LFP methodology also showed no difference between the means scores on essay 1 and essay 2, a moderately good correlation between the LFP scores on essay 1 and essay 2 (r=.665, p<.001), and much weaker, but still significant, correlations between the LFP scores and the Levels Test (for essay 1, r=.303, p<.05); for essay 2, r=.287, p<.05).

Figure 2 shows a more detailed analysis of the P_Lex data for three subjects at different

Figure 2: data for three illustrative cases, n=10 to 250 words



levels of proficiency: learner A is an advanced learner, learner B an intermediate level learner and learner C a low level learner. The figure shows the effects of using different lengths of text in the P_Lex analysis. P_Lex values were calculated for the first ten words of each text, the first 20 words of each text, the first 30 words of each text, and so on up to a maximum of 240 words. There is some variation in the P_Lex scores as the sample size increases. However, it is clear from figure 2 that P_Lex is essentially stable from about 120 words, and that texts of this length clearly discriminate between the proficiency levels illustrated. Indeed, we could argue that texts shorter than 120 words can also discriminate, but at this level the P_Lex values appear not to be reliably stable. Interestingly, low-level texts appear to stabilise faster than higher-level texts -- subject C stabilises at about 90 words, while subject A stabilises at about 150 words. This is a convenient outcome for us, as it suggests that the short texts typically extractable from low-level learners may still be long enough for evaluation purposes.

The data in this figure suggest that it might be possible to get reliable P_Lex data from texts that were considerably shorter than the texts analysed in the main experiments. We therefore repeated the analysis described in the previous section, using data from all 49 subjects, but processing only the first 150 words of each text. These data, summarised in table 6, are essentially identical to the data produced with the 250 words analysis: no difference between the means scores on the two texts, good correlations between the P_Lex scores on the two tests, and a modest correlation between the P_Lex scores and scores on the Levels Test. This suggests that P_Lex is effective even with texts lengths that are considerably shorter than the minimum figures recommended by Laufer and Nation (1995) with their LFP measure. In fact, a detailed examination of the data indicates that 120 words might be a reasonable lower bound for the analysis we have described.

Discussion

The data reported above suggests that the P_Lex methodology is basically a reliable one,

Table 6: Mean lambda scores and correlations – 150-word texts.

Levels Test	mean	51.02	Essay 1 mean	1.465	Essay 2 mean 1.295
	sd.	7.29	sd.	0.65	sd. 0.54

correlation - essay 1 and essay 2	r=0.655 p<.001
correlation - essay 1 and Levels Test	r=0.555 p<.001
correlation - essay 2 and Levels Test	r=0.371 p<.001

which produces data very similar to the data produced by LFP. However, P_Lex has the advantage that it seems to work with much shorter texts than the recommended minimum text length for LFP, and this makes it a more useful tool for analysing the output of L2 learners, particularly lower-level learners.

The question of validity is much more awkward to deal with. There are two basic problems here.

The first problem is that there are no other tests of productive vocabulary with which we can compare these data. Our approach here has been to use the so-called Productive Version of the Levels Test as a comparison point, but there are a number of reasons for viewing the Levels Test as a poor instrument when it comes to measuring productive vocabulary: it gives subjects very little freedom of choice in their responses, which are highly constrained by the context provided in the items. Nor does it allow them to display their vocabulary knowledge freely, as it tests only a very small number of items. In short, the Levels Test is constrained in a way that is totally different from P_Lex, and given these fundamental differences between the two tests, we should perhaps not be surprised that the correlations we found between P_Lex and the Levels Test were only modest.

The second problem concerns our selection of 'difficult' words. In the work reported here, we have defined 'difficult' in terms of frequency, a practise that is largely unquestioned in this field. The version P_Lex used here used Nation's (1984) word lists as a way of discriminating between 'easy' words and 'hard' words. We arbitrarily assigned words in Nation's 1000 word list to the former category, along with proper nouns, numerals and geographical derivatives, while any other words were assigned to the latter category. We think, however, that there might be a case for exploring alternative ways of characterising vocabulary. Specifically, we think that 'difficult' vocabulary is not entirely to be defined in terms of frequency: words are unusual in particular contexts for particular groups of L1 speakers, and it may not be possible to draw up a list of 'difficult' words that applies to all contexts and all L1 groups. This suggest to us that P_Lex might be most effective if it were combined with a set of standardised tasks -- e.g. picture description tasks -- where data can also be elicited from native speakers, and task-specific word lists could be constructed on the basis of these data. For example, it might be the case that native speakers asked to describe a particular picture make use of very specific vocabulary, which will be very common in this

specific context, but unusual in other contexts. Learners describing the same picture might also use 'difficult' words, but these words only really indicate good vocabulary control if they come from the same set of difficult words that the native speakers use in this context. Other 'difficult' words would then actually indicate a lack of appropriate vocabulary. A task-specific vocabulary list would be able to distinguish these cases. Some work along these lines is currently in progress in Swansea.

Conclusions

In this paper, we have outlined an alternative approach to the question of describing the vocabulary resources contained in a text. Like Laufer and Nation's (1995) Lexical Frequency Profile, P_Lex compares the lexical content of a text with external norms, essentially with frequency lists. In fact, both LFP and P_Lex use the same frequency lists, taken from Nation (1984). However, P_Lex seems to have a number of advantages over LFP in that it works well with shorter texts, and this makes it particularly suitable for use with low-level learners. P_Lex also has better measurement characteristics than LFP does -- it isn't a ratio, and it is anchored on zero -- and this makes it more amenable to standard statistical treatments than LFP.

Given these advantages, we think that P_Lex might turn out to be the kind of tool that will have a number of interesting and very practical applications. Our own work to date has largely been concerned with using P_Lex as a way of evaluating texts used in examinations. We might expect difficult examinations would use texts that had higher P_Lex values than easier examinations, and on the whole this turns out to be the case. P_Lex can thus be used to provide an objective support for examination setters' hunches about the appropriateness of a test passage for a particular group of students. Our more recent work has been more concerned with using P_Lex as a way of assessing the vocabulary resources commanded by learners at different levels of proficiency, and we think that in the longer term this methodology might be used to provide objective support for the intuitive and subjective judgements that examiners are required to make in oral assessments of speakers' abilities. At the moment, we are some way from this goal, but we think that the methodology holds more than a little promise.

Notes

experimental versions of P_Lex and a related programme K_Lex are available from our website:
http://www.lognostics.co.uk/

We would like to acknowledge the support of colleagues at Concordia University Montreal, and the support of FCAR in the development of these ideas.

References

Arnaud, PJL
the lexical richness of L2 written productions and the validity of vocabulary tests. In: T Culhane, C Klein-Braley and DK Stevenson (eds.) *Practise and Problems in Language Testing*. Department of Language and Linguistics, University of Essex, occasional papers 29. 1984.
Broeder, P, G Extra, R van Hout, S Stromqvist and K Voinmaa.
Processes in developing lexicon. Tilburg: KU Brabant. 1988.
Engber, CA
The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4(1995), 139-155*.
Herdan, G
Type Token Mathematics. The Hague: Mouton. 1960.
Laufer, B and ISP Nation
Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16(1995), 307-322.
Laufer, B and ISP Nation
A vocabulary-size test of controlled productive ability. *Language Testing 16(1999), 33-51*.
Malvern, D and B Richards
A new measure of lexical diversity. In: A Ryan and A Wray (eds.) *Evolving Models of Language*. Clevedon: Multilingual Matters. 1995.
Miller, JF and T Klee
Computational approaches to the analysis of language impairment. In: P Fletcher and B MacWhinney (eds.) *The Handbook of Child Language*. Oxford: Blackwell. 1995.
Nation, ISP
Vocabulary Lists: words, affixes and stems. Victoria University of Wellington: English Language Institute. Occasional publications 12. 1984.
Xue, Guoyi and ISP Nation
A University Word List. *Language Learning and Communication, 3*(1984), 215-229.

This paper first appeared in *Prospect* 16,3(2001), 5-19.

Appendix A: How the lambda scores are calculated

P_Lex is based on the assumption that 'hard' words are relatively rare events, and that most of the words that occur in a text are simpler, high frequency words. This means that we would expect most 10 word segments to contain only one or two hard words, and that segments containing three or four words will be relatively unusual. Distributions with these characteristics are often well described by *Poisson distributions*. The first recorded use of these distributions was a study of the number of Prussian cavalry officers kicked to death by their horses. Clearly, most days ought to contain no deaths from horse kicks, but it would not be unusual to find one Prussian cavalry officer kicked to death in a single day. However if we found a day where 10 officers received fatal kicks, then we might want to infer that something unusual happened on that day.

The advantage of fitting Poisson curves to our data is that these curves are conveniently described by a compact formula:

$$P_N = (\lambda^N * e^{-\lambda})/N!$$

The critical value in this formula is the variable λ (lambda), which defines the overall shape of the curve. If we know the value of lambda, then we know what the curve will look like, and this means that we can use lambda as a shorthand for describing data like the sample presented in table 3. These data are in fact that close approximation to a Poisson curve with a lambda value of 0.92, as we can see from the table below, and we can use at least squares method to calculate that this value is indeed the best fitting match.

N	0	1	2	3	4	5	6	7	8	9	10
actual	.40	.40	.20	0	0	0	0	0	0	0	0
Λ=0.92	.39	.36	.16	.05	.01	0	0	0	0	0	0

Inevitably, these curves are not exact fits, and P_Lex reports and Error Figure, which shows how well the data are described by the best fitting Poisson curve. At 120 words, this error figure is typically less than 5%, which is very low considering how few data points are involved. High error figures usually indicate texts that are abnormal in some way. A text that was generally very simple, for example, but contained a number of segments that contained five or six unusual words – e.g., *capture, recapture and removal statistics for estimation of demographic parameters*, where N = 7 -- would have a profile that doesn't match the standard Poisson profile, and would therefore produce a large error score. We would perhaps not want to include a text of this sort in a group analysis, and indeed, the good figures reported in the results section can be considerably improved if we eliminate subjects with error figures over 5%.

At 200 words, the mean error value falls to about 2%. This suggests that, despite its simplifications, the model is basically a good one. The close fit between the theoretical curves and the actual data suggest that, in addition to its role as a measure of vocabulary, P_Lex might be able to function as a simple diagnostic tool, identifying L2 speakers who are showing abnormalities in the way they write in English. For example, it might be possible to use P_Lex to pick out learners who over-rely on Romance cognate words in their writing.