Can P_Lex Accurately Measure Lexical Richness in the Written Production of Young Learners of EFL?

SORAYA MORENO ESPINOSA
Department of Modern Philologies
University of La Rioja

Received: 16 February 2005 accepted version: 28 March 2005
ISSN: 1697-7467

ABSTRACT: In this paper, we aim to shed light on the measurement of L2 lexical richness in the written production of primary school learners of EFL by means of a computational tool such as P_Lex (Meara & Bell, 2001). We consider it necessary to find out whether this instrument will be appropriate to analyse their embedded vocabulary, since as far as we know, this experimental tool has not been used to analyse the written production of learners with such a low proficiency level.
Key words: L2 lexical richness; EFL young learners; Computer-mediated embedded vocabulary assessment.

RESUMEN: En este trabajo pretendemos arrojar luz sobre la evaluación de la riqueza léxica de la producción escrita de aprendices de inglés como lengua extranjera en primaria a través de una herramienta informática como P_Lex (Meara & Bell, 2001). Consideramos necesario investigar si este instrumento de evaluación en fase experimental es apropiado para analizar dicho vocabulario, puesto que no parece haber sido utilizado con anterioridad para analizar la producción escrita de aprendices de inglés con un nivel de competencia en lengua extranjera tan bajo.
Palabras clave*:* Riqueza léxica; Jóvenes aprendices de inglés como lengua extranjera; Evaluación de vocabulario en redacciones mediada por ordenador

1. INTRODUCTION

P_Lex (Meara & Bell, 2001) seems to be an alternative approach to assessing the lexical complexity of short texts produced by second language learners of English. It has a passing resemblance to the Lexical Frequency Profile[1] (LFP) (Laufer & Nation, 1995), but its authors

[1] Lexical Frequency Profile (LFP) (Laufer & Nation, 1995) claims to be a reliable measure of lexical richness in written texts, which: (a) will quantify the degree to which a writer is using a varied and large vocabulary; and (b) will be used to examine the relationship between vocabulary knowledge and vocabulary use.

claim that P_Lex seems to have some advantages over the latter. Amongst other things, Meara & Bell (2001) claim that P_Lex works well with short texts, whereas the LFP requires texts over 200 tokens to obtain stable scores. As Meara (2001) notes, the results it produces need to be treated with appropriate caution, since it is not a well-tested instrument.

P_Lex may not be stable in texts shorter than 120 words, as Meara and Bell (2001) point out. However, Bell (2003) states that if the minimum text length is not reached, that does not mean that we cannot use P_Lex with short essays, but rather that our degree of confidence in the results is reduced. Bearing in mind these issues, we still feel that the present investigation is warranted, since we believe there is a need to investigate L2 young learners' productive vocabulary. Thus, we aim to test whether P_Lex can accurately measure L2 lexical richness in the written production of young learners, who are likely to produce really short texts.

There are a great variety of word-based measures, Bell (2003) divides them in two groups: (a) those which use internal evidence and usually assert the primacy of the number of types, and (b) those which consult outside the text for their evidence, whether referring to frequency lists or to comparison with other members of a group. Therefore, lexical richness can be assessed by taking into account two different benchmarks: those based on the type/token ratio (TTR) and its different variations such as Root TTR (Guiraud, 1959), Corrected TTR (Carrol, 1964), Log TTR (Herdan, 1960), Malvern-Richards-Sichel D (Malvern & Richards, 1997), Advanced TTR (Daller et al, 2003) and Guiraud Advanced (Daller et al, 2003); and those based on external standards such as frequency lists. P_Lex is included within the latter group, since it makes use of Nation (1988)'s frequency lists, and it is based on the underlying assumption that large vocabulary size is likely to lead to increased use of low-frequency words (Meara & Bell, 2001).

P_Lex divides the text into segments of 10 words each, and then counts the number of 'difficult' words in each segment. According to Meara (2001:1), P_Lex considers 'difficult words': "any word which is not found in a short list of high frequency words", which in practice means any word not included in the 1,000 most frequent English content words. Thus, P_Lex looks at the distribution of 'difficult words' in texts, and provides a simple index that indicates how likely the occurrence of these words is. The higher the P_Lex score, the bigger the vocabulary size of the learner is (see section 3.3.1. for a description of this tool).

The purpose of this study is to test whether P_Lex can actually produce accurate results, when assessing the lexical richness in a composition written by Spanish young learners of EFL, and whether it is necessary to standardise the length of the texts used for analysis. Furthermore, we aim to check whether P_Lex can discriminate between groups of learners at different levels of proficiency, by correlating its results with an independent measure of learners' proficiency; and we also consider whether there is any relationship between lexical proficiency and young learners' writing ability.

LFP makes use of a computer program which performs lexical text analysis, on the basis of different frequency levels. Thus, the VocabProfile package sorts all the words in a composition, into a four-category profile: (a) the most frequent 1,000 English words; (b) the second thousand most frequent words; (c) academic words; (d) words that are not included in neither of the previous lists.

The paper will be structured as follows: First we will briefly make reference to studies that have employed P_Lex. Secondly, we will put forward our specific objectives and present the methodology of our research. Finally, we will describe and analyse the data obtained.

2. State of the art

Different procedures have been proposed to measure the different dimensions of lexical richness, ranging from the type/token ratio and its different variations (as already stated), to other popular measures such as lexical originality (LO), lexical density (LD), lexical sophistication (LS), and lexical variation (LV) (see Laufer & Nation, 1995). Thus, there are a great variety of studies that have aimed to measure L2 lexical richness, which range from those that: (a) have proposed new measures of lexical richness such as the Lexical Frequency Profile (Laufer & Nation, 1995), P_Lex (Meara & Bell, 2001), Advanced TTR and Guiraud Advanced (Daller et al, 2003), Measure of Lexical Richness (MLR) (Vermeer, 2004); (b) measure the relationship of lexical proficiency to the quality of ESL compositions (Arnaud, 1984; Linnarud, 1986; Astika, 1993; Engber, 1995); and (c) have compared different measures of lexical richness (Laufer, 1991; Vermeer, 2000; Bell, 2003; Perkins, 1980; Mullen, 1980; Daller et al, 2003; Jarvis, 2002; Lenko, 2002; Vermeer, 2004).

However, there is a scarcity of investigations which have made use of P_Lex –an alternative approach to assess the lexical complexity of short texts produced by second language learners in English–, which confirms its experimental situation. We would like to call attention to the characteristics of these studies on P_Lex (see Figure 1 for a summary), since we have found a range of drawbacks in some of them. Amongst the different hindrances noticed, we would like to highlight the following:

(i) Meara & Bell (2001) and some of Bell (2003)'s investigations are carried out with heterogeneous samples of informants from a variety of different mother tongues, an issue which may prejudice their investigation since as Farhady (1982: 55) notes:

> "No matter what the purpose of the test may be, learner variables will definitely influence test scores in one way or another. That is, placement, selection, aptitude, proficiency, and other uses of language test scores will be sensitive to those who take the test.»

In this sense, we do agree with Farhady (1982) because the characteristics of learners from different L1s and educational backgrounds may actually influence results, and we would get much more reliable results with a rather homogeneous sample of informants such as in Miralpeix & Celaya (2002)'s study, who analysed the compositions produced by three different proficiency groups of Spanish learners of English as an L3 by means of P_Lex.

(ii) In some of Bell (2003)'s researches, certain aspects of the methodology are, we feel, open to question, such as: (a) relying on intuitive judgements to determine informants' proficiency level (see Figure 1); and (b) not determining in an accurate way the mother tongue of some of his informants (see Figure 1).

Figure 1. Summary of studies that have made use of P_Lex.

AUTHOR/S	SUBJECTS				MOST SIGNIFICANT OBJECTIVES AND/OR RESULTS
	No	L1	L2/ L3	L2/ L3 LEVEL	
Meara & Bell (2001)	49	Variety of L1 backgrounds	English	From lower-intermediate to advanced	Their analysis confirms that: (a) P_Lex is reliably stable across administrations; (b) it can discriminate different levels of proficiency; (c) there is a moderate but still significant correlation between P_Lex and the productive version of the VLT[2] and (d) it works well with much shorter texts than the minimum text length for the LFP, considering that P_Lex produces stable results from about 120 words
Miralpeix & Celaya (2002)	72	Bilingual Spanish and Catalan speakers	English	Primary (7th grade =12 year olds) and secondary education (1 BUP and COU)	They ask P_Lex for a report in each composition at 50 words to keep length constant (only the first five 10-word segments were identified and processed). Their results show that P_Lex discriminates significantly between two language levels out of three: as the level increases, the possibility of producing less frequent words increases as well.
Bell (2003)	25	Variety of L1 backgrounds	English	Intermediate to advanced	(a) P_Lex can accurately distinguish between two proficiency groups; (b) it is able to produce reliable results from two pieces of writing by the same person; and (c) 150 tokens seems to be the minimum length required for assessing lexical richness.

[2] VLT stands for Vocabulary Levels Test (Laufer & Nation, 1995).

Bell (2003)	62	Variety of L1 backgrounds	English	Variety of proficiency levels, not clearly stated	(a) The correlations between P_Lex and LFP are very good which suggests that the two measures are tracking a similar phenomenon; and (b) the simple score produced by P_Lex is a clear advantage over the four-figure result produced by LFP.
Bell (2003)	1	Japanese	English	Within the upper-intermediate to advanced range	(a) Even though P_Lex is far less sensitive to the length of the text than the Type/Token Ratio, the scores developed from very short texts are not always going to be reliable
Bell (2003)	6	Japanese or Chinese[3]	English	Two at each of three proficiency levels[4]: intermediate, upper-intermediate and advanced	(a) P_Lex scores do not remain stable across several pieces of writing by one person; (b) P-Lex is able to distinguish different writers irrespective of the number of tokens used in the sample, that is, it is not really necessary to standardise for text length; and (c) P_Lex scores do not correlate well with native-speaker judgments[5] of the overall quality of a piece of writing.
Bell (2003)	1	Korean	English	He had a reasonable[6] level of English at the beginning of the experimental period	Bell claims that: (a) P_Lex can be an effective method of tracking development of the lexicon over time; and (b) the scores from long and short texts can be compared without standardising for length, since there is a good correlation between both.

[3] The author of this investigation does not know exactly the mother tongue of his 6 informants.

[4] These global levels were not determined by a placement test but on teachers' opinion, which represents a drawback in this investigation.

[5] From our viewpoint, this investigation has some serious drawbacks on the basis of the assessors' judgement: (a) they are not given a specific scoring scale, but they are told to follow the criteria laid down for IELTS examiners; (b) judges are influenced by different factors that are not reflected in P_Lex scores such as grammatical and discursive appropriateness; and they are not given a specific rating scale on which to base their lexical ratings. Therefore, if we are not comparing like with like, and raters do not share a given rating scale, it is quite likely that there will not be a positive correlation between the two sets of results.

[6] This assessment of the subject's proficiency level seems somewhat vague, and is based on the researcher's personal intuition.

From this brief review of work using P_Lex, it emerges that our study resembles the analysis of Miralpeix & Celaya (2002) in the sense that: (a) our informants are also native speakers of Spanish within the Spanish educational system; (b) our subjects are enrolled in primary education, as some of their informants are; and (c) when averaging text length, we will use the same text length, that is, texts will be averaged to 50 words to keep length constant. It also has points in common with Bell (2003), since he also checks whether the scores from long and short texts can be compared without standardising for length, and he also matches P_Lex results against raters' judgements.

However, there are also differences between these studies and the present investigation concerning the following issues: (a) our sample of informants are young learners enrolled in 4th grade of primary education, whereas in Miralpeix and Celaya (2002)'s study the youngest subjects are enrolled in 7th grade of primary education; and in the rest of the investigations (Meara & Bell, 2001; Bell, 2003) informants are undergraduate students from a variety of mother tongues; (b) our focus of research aims to test whether P_Lex is able to provide accurate results with such young learners; (c) we aim to investigate whether accurate P_Lex scores can be achieved at full length, with a larger sample of informants than in Bell (2003)'s investigation, who has already tested that issue but with a very small number of subjects, and from different L1 backgrounds (Japanese or Chinese, and Korean); whereas our sample is larger and represents a homogeneous group of informants; and (d) we aim to correlate P_Lex results with raters' judgments based on a specific rating scale on which to base their assessment, rather than on intuition.

3. METHODOLOGY

3.1. Objectives

We aim at achieving the following specific objectives:

1. Measure Spanish 10-year-old learners' lexical richness in a composition written in English by means of P_Lex, and compare the indexes obtained when assessing texts at full length and with a standardised length of 50 tokens, in order to analyse whether it provides similar results irrespective of the number of tokens used in the composition.
2. Examine whether P_Lex is able to distinguish between groups of learners at different levels of lexical richness by correlating the results with an independent measure of proficiency.
3. Check whether there is any relationship between lexical proficiency and writing ability in L2 young learners writing, since as Engber (1995) already noted lexical proficiency influences reader judgments of the overall quality of a written composition. Thus, if EFL teachers can make use of an objective index of lexical richness provided by a tool such as P_Lex, they will have objective data on which to base their subjective judgements.

We believe that by achieving these goals, we will be able to provide new insights, especially valuable for EFL/ESL primary teachers and researchers within the Spanish educational context, which may also be extrapolated to other contexts and proficiency levels. Moreover, these results will provide a preliminary basis for a larger study of the acquisition and development of lexical competence in young Spanish learners of English, within which this investigation is subsumed.

3.2. Informants

Two hundred and seventy-one subjects have participated in this investigation; however eighty-seven of them were discarded because their compositions did not achieve the minimum length of 50 tokens. Thus, our informants comprise a sample of one hundred and eighty-four 10-year-old Spanish learners of English as a foreign language (91 boys and 93 girls) enrolled in fourth grade of primary education in four different schools[7] in La Rioja. They had received 162 hours of formal instruction in EFL, and their proficiency level reported by a placement test was rated to be beginner level (corresponding to level A1 on the basis of the Common European Framework of Reference).

3.3. Instruments and procedures

Our data gathering instrument was a composition, which was written as part of a normal class, early in the second term. At the beginning of the task, clear, general instructions were presented orally and in writing in the students' mother tongue, that is, Spanish, so as to ensure that informants were able to understand the topic they were being requested to write on. Thus, our informants had thirty minutes to write in English the following composition:

> You are going to go to England to stay with a family for a month. The family are called Mr. and Mrs. Edwards and they have two children Peter and Helen. They live in Oxford. Write a letter to them introducing yourself and telling them about your town, your school, your hobbies and any other interesting things.

Two instruments were used to measure the output of the written compositions: P_Lex (see 3.3.1) and the ESL Composition Profile (see 3.3.2), which was our independent measure of learners' proficiency.

3.3.1. P_Lex

As already mentioned, P_Lex is an exploratory tool that allows teachers and/or researchers to assess the lexical difficulty of texts. It divides the text into segments of 10 words each, and then provides a profile showing the proportion of 10-word segments containing 0 difficult words[8], the proportion containing 1 difficult word, so on and so forth, up to 10.

[7] Two of them were state schools, the other two were private schools receiving state subsidy.

[8] P_Lex assumes that difficult words are infrequent occurrences, conforming more or less to the Poisson distribution (by being strongly skewed to the left), and it calculates the theoretical Poisson curve which most closely matches the actual data produced from the text. P_Lex describes the data curves in terms of their lambda values.

It is advisable to pre-edit[9] the texts, by correcting spelling mistakes or identifying lexical phrases by connecting them with an underscore. For instance, *take_up* will be considered as a 'difficult word', whereas *take up* will be considered as two different not 'difficult words', since they are included in the 1,000 most frequent English word list.

When the text is processed, each word is compared against the contents of the dictionary files[10]. When words not included in any of the dictionary files are encountered by the computer (e.g. because they have inflectional or derivational affixes), the researcher is able to allocate them to the correct dictionary file.

Before analysing the text, several decisions are to be taken:

- Whether or not to include in the analysis Level 0 words, which consist of 28 structure words, consisting of determiners, the most common pronouns, and past forms and participles of the verbs *do*, *have* and *be*. Bell (2003) advises the user of the test to include them, since by doing so we extend the amount of data to be analysed, a serious consideration when dealing with short texts. Thus, we decided to include them in our analysis of results.
- It is also important to be consistent in the editing process, since as Bell (2003:86) emphasises: "poor procedure or makeshift decisions at the editing stage can cause anomalous results".

At the end of the analysis, P_Lex provides the following information: (a) the number of tokens in the text; (b) the number of 10-word segments identified and processed; (c) the lambda value[11] for the text; and (d) an error value[12].

3.3.2. The ESL Composition Profile

The ESL Composition Profile (Jacobs et al, 1981) is a norm-referenced criterion which focuses on the communicative aspect of language, in order to carry out integrative and holistic assessment. This instrument measures the construct of writing proficiency in English by means of five rating scales (content, organization, vocabulary, language use and mechanics), one of them being specifically devoted to the range and appropriateness of the vocabulary used by the testee, classified into four mastery levels: *excellent* to *very good*, *good* to *average*, *fair* to *poor* and *very poor*. By taking into account Read (2000)'s construct definition, the ESL Composition Profile assesses embedded vocabulary (i.e. a measure of vocabulary which

[9] The first editing task is to correct minor errors, and this is problematic, since there is no strict measure of what a lexical error is. In agreement with Laufer & Nation (1995) and Bell (2003), and we will treat as not part of the writer's lexicon only those errors which would hinder communication. In practice, as Bell (2003:86) states: "this policy usually amounts to simply correcting minor spelling errors".

[10] The dictionary files contain Nation's frequency lists, against which the computer compares the output of learners.

[11] The lambda value is a single figure that indicates how likely the occurrence of difficult words is.

[12] The error value states how close the match is between the lambda value displayed, and the Poisson distribution generated from lambda. The smaller the error value, the more satisfactory the match. It should be noted that longer texts tend to produce smaller error values than shorter texts.

forms part of the assessment of some other, larger construct), in a comprehensive way (i.e. it takes account of the whole vocabulary content of the input material), vocabulary being assessed by taking into account the context.

We took steps to achieve high degrees of validity and reliability in our manual assessment of compositions. Validity was ensured by requesting informants to write a composition by making use of vocabulary which was familiar to all of them. We aimed to increase reliability of results by providing inter and intra-reader reliability. Special attention was given to the selection of raters, who came from the same background and were knowledgeable about the *Profile*. The compositions were given to two raters in typed form, with no indication of any judgments as to their quality.

Assessors read each composition twice and conformed to the following scoring procedures:

a) First, they read each composition once to form an overall first impression of whether communication has been achieved. Based on this initial reading, they assigned a score for the associated mastery levels of *content* and *organization*.
b) Secondly, they read each composition a second time to verify their first impression by focusing on the different aspects of the composition (*vocabulary*, *language use* and *mechanics*) and score them.

All essays were rated at a rapid pace, the average time taken between 2-3 minutes per composition. Subsequently, the final score, which ranged from 34 to 100, was reached by calculating, for each composition, the average of the scores from each assessor. Generally, for the total set of scores to be valid they should fall within a range of ten points, since that would mean that both raters are interpreting and applying the same standards and criteria for assessment. Whenever scores were not within ten points of each other, a third rater was employed, and the average of the scores from the three assessors was calculated.

4. Discussion of results

In this section, we will present the results provided by P_Lex when assessing the lexical richness of the compositions written by our sample of informants. First, we will describe and analyse the results obtained using P_Lex as an assessment tool, and we will check whether it provides broadly similar results irrespective of the number of tokens used in the composition. Secondly, we will examine whether P_Lex scores correlate with the overall results of the ESL Composition Profile, by identifying weak and strong students on a similar basis. Furthermore, we will analyse the relationship between productive vocabulary and writing ability in young learners' output.

One hundred and eighty-four compositions, whose text length ranges from 50 to 311 words have been assessed by means of P_Lex. The mean number of word tokens produced by our 184 informants was 106. The assessment of compositions has been approached from two different perspectives: (a) text length has been standardised to 50 words, as in Miralpeix and Celaya (2002)'s study; (b) text length has not been standardised, but has been left at full length.

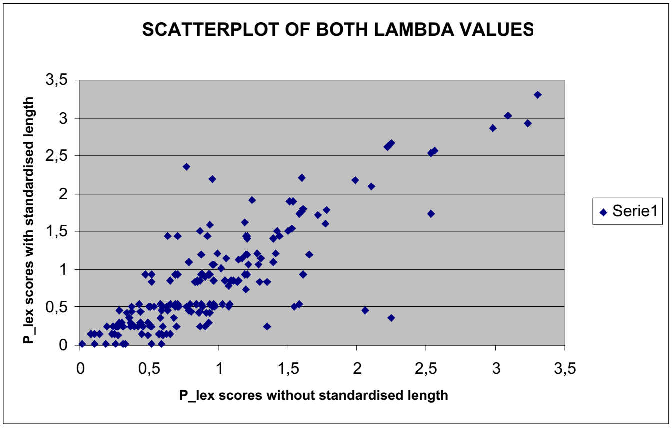
The mean lambda values reported by P_Lex with an average text length of 50 tokens, together with the standard deviation of these values can be seen in Table 1. By taking into account that according to Meara and Bell (2001), lambda values typically range from 0 to about 4.5, we can observe a value of 0.8229 indicates a low level of lexical richness, since it is closer to the minimum value than to the expected maximum value.

*Table 1. Mean lambda scores (*n *= 184).*

	Text standardised at 50 tokens	**Text length at full length**
Mean Lambda value	0.8229	0.9336
SD	0.6833	0.6091

If we approach the analysis of results by taking into account compositions at full length (see Table 1), we observe that the mean lambda value has slightly increased (mean lambda value = 0.9336), and the standard deviation is still rather high, which implies that there is a lot of variation within each group of scores that range from a minimum score of 0.02 to a maximum of 3.30 in both sets of scores. However, the positive correlation between the two sets of P_Lex scores ($r = 0.828$; $p < 0.001$) indicates a very low probability of getting these results by chance of fluctuation, and it suggests that P_Lex works for short text lengths. The actual strength of the correlation of the two sets of lambda values (with and without standardised average length) is reflected in the following scatter plot (see Figure 2).

Figure 2. Scatter plot of both sets of lambda values (with and without standardised length).



By taking into account the frequency distribution of results with and without standardised length (see Figures 3 and 4), we can see that the scores are heavily skewed to the lower end of the range, but are relatively widely spread out. In Table 3, we can observe how the percentage of informants decreases as the lambda value increases from 1 point onwards at full length, whereas we obtain constant decreasingly scalable scores with a standardised text length at 50 tokens (see Table 3). The results indicate that more students received a lower lambda value than a higher one. The issue of whether those learners that were awarded with a higher lambda value were more proficient that those who were given a lower one, will be addressed later on in this paper, by checking the lambda values against the ESL Composition Profile, as an independent measure of learners' proficiency.

Table 3. Frequency distribution of scores.

	PERCENTAGE OF INFORMANTS	
LAMBDA VALUES	AT FULL LENGTH	AT 50 TOKENS
0.00 – 0.49	24.46%	37.50%
0.50 – 0.99	41.30%	32.61%
1.00 – 1.49	19.02%	14.13%
1.50 – 1.99	8.70%	8.70%
2.00 – 2.49	2.72%	2.72%
2.50 – 2.99	2.17%	3.26%
3.00 – 3.49	1.63%	1.09%

Figure 3. Frequency distribution of P_Lex scores at full length tokens.

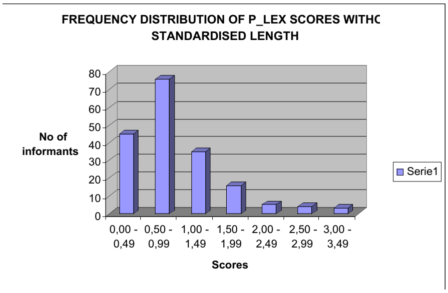
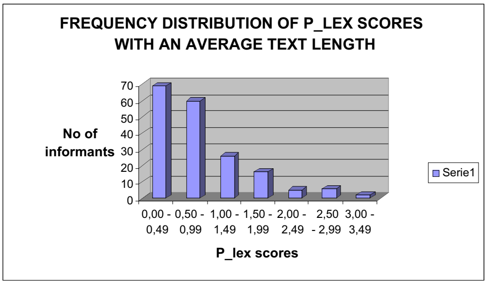


Figure 4. Frequency distribution of P_Lex scores standardised at 50.



A related samples t-test was done on the two sets of scores, pairing the two scores for each text (standardised, non-standardised), giving a highly significant value of t ($t = 3.88$; $p < 0.001$), which shows that there are very significant differences between the two sets of scores and therefore it suggests that the longer texts produce higher scores, which contradicts the issue of P_Lex working well with short texts. Therefore, it seems that the scores for the full texts reveal a higher lexical richness than the standardised texts.

In order to address our second goal, we used the results of the Vocabulary band in the ESL Composition Profile to divide the subjects into two sub-groups: Group A (*n* = 129) comprised the subjects that had achieved a score between 7 and 13 and whose embedded vocabulary was measured to be *very poor* to *fair;* and group B (*n* = 55) comprised those informants that had been awarded a score in the range 14 to 20, rated to be *average* to *excellent*.

The mean values of both groups of learners reported by the ESL Composition Profile, for the vocabulary band and the whole composite score, together with the standard deviation of these values can be seen in Table 4.

Table 4. ESL Composition Profile results.

		GROUP A (*n* = 129)	**GROUP B (*n* = 55)**
VOCABULARY BAND	**MEAN**	11.49	15.02
	SD	1.66	1.13
COMPOSITE SCORE	**MEAN**	57.92	72.89
	SD	7.63	5.89

Independent t-tests indicated that the ESL Composition Profile scores of these two groups were reliably different (t = 12.989, df = 182, p < .001 for the total composite score; t = 14.369, df = 182, p < .001 for the vocabulary band of the ESL Composition Profile). We correlated the results of the vocabulary band and its overall results in the ESL Composition Profile to verify whether there was a positive correlation between vocabulary and writing ability. Results indicate that there is a positive correlation between vocabulary and overall composite scores not only in group A (r = .903, p < .001), but also in group B (r = .878, p < .001), which indicates a relationship between L2 lexical proficiency and EFL young learners' writing ability, whether we consider total composite scores or the vocabulary band in the ESL Composition Profile.

Once we checked that our two groups were reliably different and there was a positive correlation between their vocabulary and writing ability, measured by an independent criterion, we examined whether P_Lex was able to distinguish between weak and strong learners in a similar way to our independent measure.

Our results indicate that, in group A, there was a significant correlation between P_Lex scores at a standardised length of 50 tokens and the ESL Composition Profile (r = .175, p < .05), and P_Lex scores at full length and the ESL Composition Profile (r = .198, p < .05). However, we found a negative correlation in group B, between the Profile and P_Lex scores at a standardised length of 50 tokens (r = -.072, p = .601), and between the Profile and P_Lex scores at full length (r = -.038, p = .782), which suggests that P_Lex and ESL Composition Profile correlate for weaker students but not for stronger ones.

However, two independent t-tests indicate that none of the groups (A or B) are reliably different on the basis of mean P_Lex scores (t = -.520, df = 182, p = .603) at full length; and at a standardised length of 50 tokens (t = -1.805, df = 182, p = .073), which suggests that P_Lex –despite being less susceptible than other measures to the problems caused by short texts- is not effective for distinguishing different proficiency levels when assessing L2 young learners.

5. Conclusion

This investigation stemmed from our interest in discovering whether P_Lex is an appropriate tool to measure the L2 lexical richness of young learners. We aimed to overcome some of the shortcomings of previous studies, specifically by gathering compositions written by a homogeneous sample of informants.

From our viewpoint and according to the results achieved, we believe that empirical evidence suggests that P_Lex cannot appropriately be used to measure the L2 lexical richness of young learners in written compositions. Our results indicate that despite the fact that there is a positive correlation between P_Lex scores at full length and at a standardised length of 50 tokens, a t-test has shown that the longer texts produce higher scores, which contradicts the issue of P_Lex working well with short texts, probably due to the instability of scores of texts shorter than 120 words.

A positive correlation between the vocabulary band of the ESL Composition Profile and its overall scores has demonstrated that there is a relationship between lexical proficiency and writing ability in young learners, issue which scholars such as Engber (1995) had already highlighted when analysing the compositions of learners whose language proficiency ranged from intermediate to advanced.

When examining whether P_Lex was able to distinguish between groups of learners at different levels of lexical richness by correlating the results with the ESL Composition Profile, -as an independent measure of proficiency-, findings suggest that P_Lex is unable to do it, since despite the fact that the results from weaker learners show a moderate significant correlation between P_Lex results and the Profile's ones, an independent t-test indicates that neither group A nor B are significantly different on the basis of P_Lex mean results at full length and at standardised length.

Despite the discouraging results when using P_Lex to assess L2 young learners' lexical richness and its inability to distinguish different levels of EFL young learners' language proficiency, we believe that computational assessment instruments are of paramount importance for ESL/EFL teachers and researchers, since such tools may provide an objective index on which to base teachers' subjective judgements on vocabulary assessment, which in turn seems to be directly related to writing ability.

We are aware of the fact that vocabulary knowledge is many faceted (Laufer, 2001), that "no one particular measure of lexical richness can serve all purposes" (Bell, 2003:220) and that "one measure of lexis alone cannot describe the lexical texture of a composition" (Linnarud, 1986:47), therefore we believe that further research should be carried out, specifically to check : (a) whether there is a positive correlation between the results provided by P_Lex and any other measures of lexical richness; (b) whether we are able to monitor progress in the development of the vocabulary of Spanish primary learners of EFL through P_Lex results; and (c) whether P_Lex is an appropriate assessing instrument to measure the lexical richness of learners enrolled in secondary education.

ACKNOWLEDGMENTS

This investigation was supported by the Spanish Ministry of Science and Technology (Grant no. BFF2003-04009-CO2-02) as part of the research project "The development of lexical competence in the acquisition of L2 English in primary education", University of La Rioja, and Comunidad Autónoma de La Rioja (Grant F.P.I. 2005).

The authors would like to gratefully acknowledge Huw Bell's aid in providing us with a copy of his thesis. We would also like to acknowledge the insightful comments of Professor Christopher Butler, and also of three of his colleagues, on previous drafts of this paper. For any shortcomings that may remain the authors are, of course, solely responsible.

REFERENCES

Arnaud, P. J. L. (1984). "The lexical richness of L2 written productions and the validity of vocabulary tests", in *University of Essex Occasional Papers*, 29: 14-28.

Astika, G. G. (1993). "Analytical Assessments of Foreign Students Writing", in *RELC Journal*, 24, 1: 61-72.

Bell, H. (2003). *Using Frequency Lists to Assess L2 Texts*. University of Wales Swansea: Unpublished Thesis.

Carrol, J. B. (1964). *Language and Thought*. New Jersey: Prentice-Hall.

Daller, H., Van Hout, R. and Treffers-Daller, J. (2003). "Lexical Richness in the Spontaneous Speech of Bilinguals", in *Applied Linguistics,* 24, 2: 197-222.

Engber, C. A. (1995). «The Relationship of lexical proficiency to the quality of ESL compositions», in *Journal of Second Language Writing*, 4, 2: 139-155.

Farhady, H. (1982). "Measures of Language Proficiency from the Learners' perspective", in *TESOL Quarterly,* 16, 1: 43-59.

Guiraud, P. (1959). *Les charactères statistiques du vocabulaire*. Paris : Presses universitaires de France.

Herdan, G. (1960). *Type-token mathematics*. Gravenhage: Mouton.

Jacobs, H., Zinkgraf, S., Wormuth, D. R., Hartfiel, V. F, Hughey, J. B. (1981). *English Composition Program*. Rowley Mass: Newbury House Publishers Inc.

Jarvis, S. (2002). "Short texts, best-fitting curves and new measures of lexical diversity", in *Language Testing,* 19, 1: 57-84.

Laufer, B. (1991). "The Development of L2 Lexis in the Expression of the Advanced Learner", in *The Modern Language Journal,* 75, 4: 440-448.

Laufer, B. (2001). "Quantitative evaluation of vocabulary: How it can be done and what it is good for", in C. Elder et al (eds.), *Experimenting with uncertainty*. Cambridge: Cambridge University Press, 241-250.

Laufer, B. & Nation, P. (1995). "Vocabulary size and use: Lexical richness in L2 written production", in *Applied Linguistics,* 16, 3: 255-271.

Laufer, B. & Paribakht, T. S. (1998). "The Relationship Between Passive and Active Vocabularies: Effects of Language Learning Context", in *Language Learning,* 48, 3: 365-391.

Lenko-Szymanska, A. (2002). "How to trace the growth in learners' active vocabulary: a corpus-based study", in B. Ketteman and G. Marko (eds.) *Teaching and Learning by doing corpus analysis*. Amsterdam: Rodopi, 217-230.

Linnarud, M. (1986). *Lexis in Composition: a performance analysis of Swedish learners' written English*. Malmö: Liber Forlog.

Malvern, D. D. and Richards, B. J. (1997). "A new measure of lexical diversity", in A Ryan and A Wray (eds.), *Evolving Models of Language*. Clevedon: Multilingual Matters.

Meara, P. and Bell, H. (2001). "P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts", in *Prospect,* 16, 3: 323-337.

Meara, P. (2001). *PLEX v1.1*. University of Wales Swansea. Available at http://www.swan.ac.uk/cals/calsres

Miralpeix, I. and Celaya, M.L. (2002). "The Use of P_Lex to Assess Lexical Richness in Compositions Written by Learners of English as an L3", in I. Palacios (ed.), *Actas del XXVI Congreso de AEDEA. Santiago de Compostela/ University of Santiago de Compostela Press*, 399-406.

Mullen, K. A. (1980). "Evaluating Writing Proficiency in ESL", in J.W Oller and K. Perkins (eds.): *Research in language testing*. Rowley: Newbury House.

Nation, I.S.P. (1988). *Word Lists*. Victoria: University of Wellington Press.

Perkins, K. (1980). "Using objective methods of attained proficiency to discriminate among holistic evaluation", in *TESOL Quarterly,* 14: 61-69.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Vermeer, A. (2000). "Coming to grips with lexical richness in spontaneous speech data", in *Language Testing*, 17: 165-83.

Vermeer, A. (2004). "The relation between lexical richness and vocabulary size in Dutch L1 and L2 children", in P. Bogaards and B. Laufer (eds.): *Vocabulary in a Second Language*. Amsterdam: John Benjamins Publishing Company, 173-189.