Malaysian Journal Of ELT Research

ISSN: 1511-8002

Vol. 6, 2010

What Electronic Tools Can EFL Teachers Use to Assess L2 Embedded Vocabulary Objectively?

Soraya Moreno Espinosa & Mª Pilar Agustín Llach

E.O.I. "El Fuero de Logroño" & University of La Rioja

Spain

Abstract

Measuring vocabulary is of paramount importance for EFL teachers and researchers, as nowadays it is widely acknowledged that L2 vocabulary is central to the learning and acquisition of foreign languages. It is a difficult task to assess the vocabulary produced by learners in written tasks, as it implies the subjectivity of the rater and/or raters. However, EFL teachers can use a wide range of computer-mediated assessing instruments in order to analyse such vocabulary in a more objective way. In this article, we will describe and analyse three electronic instruments: the Lexical Frequency Profile (Laufer & Nation 1995), P_Lex (Meara & Bell 2001) and V_Size (Meara & Miralpeix 2007). These computer-mediated text analysers represent a sample of instruments that EFL teachers can use to assess the L2 vocabulary produced by learners in different writing tasks. We consider that our description and analysis of such assessing tools will enable them to decide which one best fits their pedagogical and/or assessment goals.

KEYWORDS: L2 vocabulary, assessment, computerised assessing instrument

Introduction

Although vocabulary used to be a neglected aspect of foreign language teaching (Meara 1980); nowadays, it is widely acknowledged that L2 vocabulary is central to the learning and acquisition of foreign languages, and it is a core part in the process of communication (McCarthy 1990; Vermeer 1992).

Not only is vocabulary important for effective communication, but it can also be related to other aspects of linguistic ability such as: (a) reading (e.g. Agustín & Terrazas 2009; Laufer 1992, 1997; Nation & Coady 1988; Nation 2006;); (b) writing (e.g. Astika 1993; Engber 1995; Linnarud 1986; Moreno *et al.* 2005); (c) listening comprehension (e.g. Nation 2006; Webb & Rodgers 2009); and even (d) general language proficiency (e.g. Alderson 2005; Cobb 2000; Meara 1996).

Thus, nowadays the importance of the relationship of L2 vocabulary to other language skills is reflected in the reorientation carried out not only in teaching and research, but also in assessment. Furthermore, the aim of L2 vocabulary assessment is not only to make decisions about what test takers have learnt in a teaching/learning context or to diagnose learners' needs, but also to progress in our understanding of the processes of

vocabulary acquisition by analysing what stage of L2 vocabulary development learners are at (Nation 1990; Read 2000). However there is still a need to find out: (a) what vocabulary test is the most suitable, as vocabulary knowledge is many-faceted and there is no single test able to measure all forms of vocabulary knowledge (Bogaards 2000; Laufer 2001; Nation 2001; Schmitt *et al.* 2001); and (b) what instruments are the most adequate to assess this language component in compositions, as the existing computational tools may pose problems such as the variability of results due to text-length, amongst others.

The literature on L2 vocabulary assessment suggests that there are no perfect vocabulary measures but there are a wide range of tools available (e.g. see Jiménez & Moreno (2005) for a review of vocabulary tests to investigate vocabulary knowledge in primary and secondary education). We come across different procedures which have been used to assess the different dimensions of lexical richness. Such measures range from the type/token ratio and its different variations, to other popular ones such as lexical originality, lexical density, lexical sophistication, and lexical variation (see Laufer & Nation 1995).

Previous research has suggested that there is a significant relationship between lexical proficiency and the holistic scores assigned by human raters (e.g. Engber 1995; Daller & Phelan 2007; Moreno *et al.* 2005). Nowadays, assessing EFL learners' essays plays an important role in high-stakes examinations. Thus, EFL learners are usually asked to write compositions on different topics in different public exams. Although such work has many strengths, since raters perform a thorough qualitative assessment of learners' writing; it also has some weaknesses. It is widely acknowledged that such assessment may rely on subjective judgements, and to provide inter- and intra-rater reliability at least two raters are required, which is time-consuming and costly (Jacobs *et al.* 1981). Therefore, as Meara & Bell (2001) point out, this could be one of the reasons to use objective measures of the lexical characteristics of L2 compositions.

In the last few years, there has been a fresh impetus on the use of computers in language assessment, being the largest move to computer-based testing the introduction of the computer-based Test of English as a Foreign Language (TOEFL) in 1998. However, computer mediated assessment has some limitations. For instance it just focuses on orthographic words rather than on lexical items (Read 2000), and Automated Essay Scorers can still be tricked to award higher or lower than deserved scores (Powers *et al.*

2002). Nevertheless, in spite of the possible drawbacks of computer-mediated assessment, research has suggested that automated essay scores can be used as a second rater in the assessment of compositions, and they provide a higher reliability than if multiple human raters are used (Attali & Burstein 2006).

In our view, computer-mediated assessing instruments offer an interesting way to tackle raters' subjectivity and time-consuming effort. Bearing in mind that there tends to be a relationship between lexical proficiency and holistic scores, we believe that the objective results obtained by means of computer-implemented vocabulary analysers may enable a single rater to provide a valid and reliable scoring of learners' compositions in EFL school contexts. Thus, the general goal of this article is to describe and analyse different electronic instruments that EFL teachers can use to assess the L2 vocabulary used by learners in different writing tasks.

The article is divided into three different parts. First of all, we consider essential to define basic tenets such as 'what a word is', because depending on its definition, we may obtain different results on the basis of its assessment (Daller *et al.* 2007). Secondly, we will describe a sample of three computer-mediated assessing instruments that EFL teachers

can use. And thirdly, we draw our conclusions, and the main pedagogical implications of our paper.

Defining Our Unit of Measurement

Although everybody knows what a word is; in practice, it is quite difficult to define it (Carter 1998; Read 2000). Scholars tend to adapt its definition to their needs using the general term *word*, without being systematic in their nomenclature. Hence, depending on their goals they may consider *types*, *tokens*, *lemmas* or *word families* as words (Daller *et al*. 2007).

If our general objective is to depict different assessing instruments, it is obvious that we should start by defining our unit of assessment. Hence, this section will be devoted to provide an overview of different definitions of what a word is, and specifying the constraints of our unit of measurement.

Words can be classified into two different categories: grammatical and lexical words. *Grammatical or function words* belong to a closed class which comprises pronouns,

articles, auxiliary verbs, prepositions, and conjunctions. These are words that hold little meaning and whose primary function is contributing to the grammatical structure of language. On the other hand, *lexical or content words* constitute an open class which includes nouns, adjectives, verbs, and adverbs. The latter are usually included in vocabulary tests, whereas the former are tested in grammar tests. One of the advantages of assessing vocabulary in written compositions is the fact of assessing both types of words.

In the assessment of vocabulary knowledge and use, there are two essential issues to be addressed: (i) *What is considered to be a word*: For instance, whether proper nouns and numbers are to be counted as words, or whether *father* and *father's* are one or two different words, amongst other things; and (ii) *What is counted as a word*: For example, are *happy*, *happiness,* and *unhappy* one or three different words? (Nation 2001).

With regard to the first issue, we come across the definition of *orthographic words*, which represent any sequence of letters (which may have a limited number of other characteristics such as a hyphen or an apostrophe) bounded on either side by a space or punctuation mark (Carter 1998). Hence, following this definition of word *won't* will be

considered as one word, and *will not* as two different words. From our point of view, the notion of orthographic word is opposed to the concept of *lexical phrases* (Nattinger & DeCarrico 1992) that subsumes words that operate as a single unit. For instance, *so to speak*, *once upon a time*, so on, and so forth.

Computer-mediated assessing instruments tend to adopt the definition of orthographic word (Bogaards 2001), as their assessment is based on frequency lists, and multi-word items do not seem to be subsumed in such lists. Another reason may be due to the fact that lexical phrases are an open-ended set of items, and therefore are more difficult to analyse manually or by computer (Read 2000).

With regard to 'what is counted as a word', in L2 vocabulary assessment, the term *word* as unit of analysis can refer to:

(a) *Tokens* or *running words,* which refer to the total number of word forms in a text (Nation 2001; Read 2000). For example, the sentence "The boy likes the food that his mum cooks" contains nine tokens.

(b) *Types*, which are the total number of different word forms in a text (Read 2000). For instance, the aforementioned sentence has eight different words or types, because the article *the* is a single type mentioned twice.

(c) *Lemmas,* which include the base word and its inflections, in which neither the word meaning, nor the word class of the base is changed (Read 2000; Nation & Waring 1997). For instance, *play*, *plays* and *played* represent one *lemma*.

(d) W*ord families* include inflected and regularly derived forms of a base word that share the same meaning (Bauer & Nation 1993; Read 1988). As Hirsh and Nation (1992, p. 692) note: "The idea behind a word family is that inflected and regularly derived forms of a known base word can also be considered as known words if the learners are familiar with the affixes".

Although there are different scholars that support this definition of word family, assuming that "if one knows the base word, little if any additional learning is required in order to understand its various inflectional and derived forms" (Read 1988, p. 14). There is an alternative trend that suggests that such assumption is not as clear-cut as it might seem, because knowing a word does not automatically imply knowing all its derivatives

and inflections (Beglar & Hunt 1999; Bogaards 2001; Jiménez & Mancebo 2008; Schmitt & Meara 1997; Schmitt & Zimmerman 2002; Vermeer 2004).

We do agree with Schmitt and Zimmerman (2002), when they point out that although it may be true that inflectional members of a word family are relatively easy to learn, derivates can sometimes be opaque and inconsistent, carrying a different learning burden. Their results yield support for the view that learners acquire partial word family knowledge, as their informants tended to know some members of the word family, mostly nouns and verbs; which suggests that especially low-level learners may be unlikely to be fully aware of all the members of a word family.

Furthermore, Nation (2001, p. 8) puts forward the notion of learning burden: "Should the irregular forms be counted as a part of the same lemma as their base word or should they be put into separate lemmas?". That is to say, are words such as *mice, brought, beaten* and *best* to be included as part of the same lemma? Obviously knowing an irregular past tense such as *brought*, does not require the same learning burden as learning a regular past tense such as *played*. Unfortunately, there is no systematic approach on these issues, and scholars just state the criteria they have followed.

With regard to *word families*, Bogaards (2001) points out that there are studies in which its definition takes semantic differences into account (e.g. Nagy & Anderson 1984); whereas in other studies (e.g. Laufer & Nation 1995), they do not discriminate between different families having the same name, i.e. homographs, words that are polysemous, and cover more than one lexical unit.

Computerised assessing instruments usually analyse words to group them into lemmas and/or word families. However, the opposite procedure can also be carried out by *Familizer*, an on-line computer programme which can expand a raw word list into a word family (available from http://www.lextutor.ca) (Cobb 2000). Unfortunately, *Familizer* does not discriminate between different families having the same word form. From our point of view, this programme needs some further refinement as it does not include the irregular comparatives of common adjectives as *good* or *bad*, despite including irregular verb forms of verbs such as *sing*.

According to Nation (2007, p. 39), the unit of analysis "should match the use to which the data is put". Thus, on the basis of previous studies, and to avoid overestimation or

underestimation of vocabulary knowledge, he suggests that if we are to analyse learners' productive vocabulary, "the lemma is the most valid unit of counting" (p. 39), whereas for receptive uses, "the word family is a more valid unit" (p. 39).

Furthermore, we also need to define what knowing a word means. It is generally acknowledged that word knowledge is multi-faceted (Daller *et al.* 2007; Nation 2001; Read 2000; Schmitt 1998). As Nation (2001, p. 23) points out "there are many things to know about any particular word and there are many degrees of knowing".

Many attempts have been made to define 'what knowing a word means' (e.g. Bogaards 2000; Laufer 1997; Meara 1996; Nation 1990, 2001; Richards 1976). In the last few years, Nation's (2001) definition of what is involved in knowing a word has been found amongst the most influential (Read 2004). Such definition involves receptive and productive knowledge of its *form* (pronunciation, spelling and word parts), *meaning* (referring to the form and meaning link, concepts and referents and word associations), and *use* (grammatical functions, collocations and constraints on use). Table 1 summarizes the components of productive word knowledge.

Table 1 What is involved in knowing a word productively? (Adapted from Nation 2001, p. 27)

WHAT IS INVOLVED IN PRODUCTIVE WORD KNOWLEDGE?		
FORM	**SPOKEN**	How is the word pronounced?
	WRITTEN	How is the word written and spelled?
	WORD PARTS	What word parts are needed to express meaning?
MEANING	**FORM AND MEANING**	What word form can be used to express this meaning?
	CONCEPTS AND REFERENTS	What items can the concept refer to?
	ASSOCIATIONS	What other words could we use instead of this one?
USE	**GRAMMATICAL FUNCTIONS**	In what patterns must we use this word?
	COLLOCATIONS	What words or types of words must we use with this one?

CONSTRAINTS ON USE	Where, when and how often can we use this word?

When testing vocabulary knowledge, there is no single test able to tap into all the different aspects (Bogaards 2000; Schmitt *et al.* 2001), and Laufer (1998, 2001) proposes a 'multiple test approach' to test the different aspects of vocabulary knowledge. Nation (2001) provides a useful table for deciding what aspects of vocabulary knowledge to test (see Table 2). He also points out that (2001, p. 362): "When testing vocabulary, it is important to distinguish between how well a word is known and how well a word is used". In this paper, our computer-mediated assessing instruments address the latter, as they aim at assessing some aspects of productive vocabulary use within the more general framework of communicative competence. Hence, L2 embedded vocabulary assessment can be said to be subsumed within an interectionalist perspective (Read & Chapelle 2001).

Table 2 Aspects of productive word knowledge for testing (Adapted from Nation 2001, p. 347)

WHAT ASPECTS OF PRODUCTIVE WORD KNOWLEDGE ARE WE TESTING?

FORM	**SPOKEN**	Can the learner pronounce the word correctly?
	WRITTEN	Can the learner spell and write the word?
	WORD PARTS	Can the learner produce appropriate inflected and derived forms of the word?
MEANING	**FORM AND MEANING**	Can the learner produce the appropriate word form to express this meaning?
	CONCEPTS AND REFERENTS	Can the learner use the word to refer to a range of items?
	ASSOCIATIONS	Can the learner recall this word when presented with related ideas?
USE	**GRAMMATICAL FUNCTIONS**	Can the learner use this word in the correct grammatical patterns?
	COLLOCATIONS	Can the learner produce the word with appropriate collocations?

CONSTRAINTS ON USE	Can the learner use the word at appropriate times?

In this section, we have aimed at reviewing different approaches to what a word and word knowledge is. As far as our battery of electronic assessing instruments are concerned, our definition of 'what is considered to be a word' will be based on orthographic words, rather than on words as lexical phrases; as they do not seem to be included in the frequency counts, which our sample of computer-mediated instruments use. Furthermore, we should also take account of a general framework that may subsume the wide range of assessment procedures. Following Read (2000), vocabulary assessment can be divided into three different dimensions. Hence it can be assessed: (a) either as discrete, or as an embedded element within a larger construct; (b) in a selective or comprehensive way; and (c) as a context-dependent or as a context independent element (see Table 3 for a definition of the different dimensions).

Table 3 Dimension of vocabulary assessment (Read 2000, p. 9)

Discrete	**Embedded**
A measure of vocabulary knowledge or use as an independent construct	A measure of vocabulary which forms part of the assessment of some other, larger construct
Selective	**Comprehensive**
A measure in which specific vocabulary items are the	A measure which takes account of the whole vocabulary content of the

This article aims at reviewing a sample of electronic measures that assess L2 embedded, comprehensive, and context dependent vocabulary. In the following section, we will describe a sample of the best-known computer-mediated vocabulary analysers so that EFL teachers may analyse which one could best fit their assessment and/or pedagogical purposes.

Electronic Tools that EFL Teachers Can Use to Assess L2 Embedded Vocabulary Objectively

As far as measures of learner production are concerned, we can use lexical statistics to analyse their use of vocabulary. The general term used for those lexical measures is *lexical richness* (Read 2000). Following Read (2000), we consider that lexical richness is a general quality of good writing. To measure writing quality, different measures based on lexical statistics are found in the literature. Meara & Bell (2001) divide them into: (a) *intrinsic measures of lexical variety*, which refer to the types and tokens that are in the text, without making reference to any external criteria; and (b) *extrinsic measures of lexical richness*, which consult outside the text with additional information about the words being used, whether referring to frequency lists or to comparison with other members of a group.

The purpose of this section is to review a sample of the most widely used comprehensive vocabulary measures, which are suitable for the assessment of vocabulary within a larger construct such as a written composition (Read 2000). Our focus of attention will be extrinsic measures, which usually employ frequency-based wordlists as a yardstick to compare the productive vocabulary of test takers.

In the literature, we find a wide amalgam of such measures, for instance the Lexical Frequency Profile (Laufer & Nation 1995), P_Lex (Meara & Bell 2001), WordClassifier

(Goethals 2005), ADELEX Analyser (http://www.ugr.es/~inped/ada/), Jacet 8000 Level Marker (http://www01.tcp-ip.or.jp/~shin/J8LevelMarker/j8lm.cgi), Frequency Level Checker (http://language.tiu.ac.jp/flc/tool.html), and V_Size (Meara & Miralpeix 2007), amongst others. Describing all the extrinsic measures would go beyond the scope of this paper, therefore we aim at depicting a sample of the most widely employed ones, including the Lexical Frequency Profile (Laufer & Nation 1995), P_Lex (Meara & Bell 2001), and V_Size (Meara & Miralpeix 2007).

Lexical frequency profile

The *Lexical Frequency Profile* (LFP) has been widely used to analyse L2 vocabulary use in free writing (e.g. Laufer 1998; Laufer & Nation 1995; Lenko 2002; Meara & Bell 2001; Muncie 2002). According to Laufer and Goldstein (2004), the LFP is an indirect test of meaning, in which it is shown the proportion of frequent versus infrequent correct form-meaning links.

The LFP makes use of a computer programme (i.e. *VocabProfile*) which performs lexical text analysis, on the basis of different frequency levels. The programme can be downloaded free of charge from http://www.vuw.ac.nz/lals/. It can be used to compare

the vocabulary of up to 32 different texts at the same time, by facilitating information according to the series of options set. Thus, it can provide a distribution figure, a headword frequency figure, a family frequency figure and a frequency figure for each of the texts the word occurs in. It can also be used to find the coverage of a text by certain word lists, to discover shared and unique vocabulary in several texts and even create word lists based on frequency and range (Nation 2005). The Web version of this programme has been developed by Tom Cobb (available from http://www.lextutor.ca). Although as he notes, it does not handle extremely large texts as the off-line programme does.

Thus, texts are typed in, without lemmatisation and saved as files in ASCII format (i.e. text format with line breaks). Subsequently, the VocabProfile package sorts all the words in the file, into a four-category profile. Although a word is defined by the programme as a base form with its inflected and derived forms, i.e. a word family; it also provides a lexical frequency profile on the word types and tokens of the text. And it shows the absolute and relative proportion of words in compositions covered by the frequency lists. Hence, it classifies words into four categories: (a) the most frequent 1,000 English words;

(b) the second thousand most frequent words; (c) academic words (AWL); and (d) words that are not included in either of the previous lists (NIL).

A minimum text length of 200 tokens is required to get stable results. Before running the programme, L2 texts are advised to be edited in the following way: Spelling errors that do not distort the word are corrected in order to make the word recognisable by the computer; whereas proper nouns, incorrectly used words and semantically incorrect words are omitted, since they cannot be considered as part of the subject's productive vocabulary (Laufer & Nation 1995).

However, scholars like Coniam (1999) do not agree with Laufer and Nation's (1995) editing proposal, and he suggests that learners should only be given credit for full word knowledge, rather than for partial word knowledge. Furthermore, he suggests that the word-list family concept is not a very consistent procedure as learners may get credit when it may not be due. For instance, if a learner knows the word *awful*, it does not mean that he/she knows the word *awe*. We do agree with him and following Nation (2007), we consider that the lemma should be analysed to avoid overestimation of productive vocabulary knowledge. As regards vocabulary knowledge, Meara (1996) points out that

the basic dimension of lexical competence is vocabulary size; therefore, learners with big vocabularies are more proficient in different language skills than learners with a smaller vocabulary size.

Laufer and Nation (1995) suggest different LFP measures on the basis of learners' proficiency (see Table 4). Thus, for less proficient test takers, a distinction should be made between the first 1,000 most frequent words, the second 1,000, and any other vocabulary. Whereas for advanced students, the profile could distinguish between the second 1,000 most frequent words, words included in the AWL, and words that are not included in any of the previous lists.

Lenko (2002) proposes the Condensed LFP for advanced learners, which replaces the four-figure profile computed by the Standard LFP with two bands: the percentage of words in a text belonging to the first two frequency bands, and a percentage of words beyond the 2,000 level. According to Lenko (2002), the Condensed LFP is better presented as the proportion of highly frequent words (i.e. up to the 2,000 most frequent words) and infrequent words (i.e. words beyond the 2,000).

Table 4 Possible LFP's measures

Standard LFP (Laufer and Nation 1995)	**LFP for less proficient learners** (Laufer and Nation 1995)	**LFP for advanced learners** (Laufer and Nation 1995)	**Condensed LFP for advanced learners** (Lenko 2002)
1,000	1,000	2,000	1,000 + 2,000
2,000	2,000	AWL	
AWL	AWL +NIL	NIL	AWL +NIL
Not in the lists (NIL)			

Although Laufer and Nation (1995) claim that LFP shows a number of advantages which are not present in other measures of lexical richness, it has also received criticisms. Meara and Fitzpatrick (2000) claim that a free productive vocabulary test such as LFP is problematic since it is context limited, in the sense that it is unclear whether the tasks

LFP makes use of, do really encourage testees to display a rich variety of vocabulary. They also note that a huge amount of text is required so that non-native speakers will be able to elicit some infrequent words, issue which can represent a drawback. Furthermore, Meara and Bell (2001) note that it may not discriminate well between the texts produced by low-level learners, as they tend to produce only very few low frequency words, which may not be enough to show distinguishing lexical frequency profiles.

The validity and reliability of the tool were demonstrated in Laufer and Nation's (1995) study, in which they claimed that: (a) the LFP was able to discriminate between undergraduate learners at different proficiency levels; (b) it was topic independent, that is to say, it provided stable results for two compositions by the same learner; and (c) it correlated well with an independent measure of vocabulary knowledge, the productive version of the *Vocabulary Levels Test*.

In 2005, Meara questioned the validity and reliability of the tool through different computer simulations, and he suggested that the claims made by Laufer and Nation "may be less robust than they made out to be" (p. 46). Meara (2005) points out that: (a) Laufer and Nation's (1995) claim that the LFP provides similar stable results across two learner

compositions written by the same subject is a very weak claim, being "a null hypothesis, with a very high probability of being confirmed by chance data (p. 44)"; (b) his computer simulations suggest that the LFP does not reliably distinguish between groups of learners at different levels of proficiency, and probably does not produce strong correlations between the LFP of different texts produced by the same learner.

Despite the possible drawbacks of this assessing instrument, we still believe that analyzing the lexical profile of learners' texts may be useful for EFL teachers to know learners' weaknesses and strengths, and especially what stage of L2 vocabulary development learners are at (Nation 1990).

P_Lex

P_Lex v2.0 (Meara & Bell 2001) is an exploratory tool that allows teachers and/or researchers to assess the lexical difficulty of texts. However, as Meara (2001) notes, the results it produces need to be treated with appropriate caution, since it is not a well-tested

instrument. However, it seems to be an alternative approach to assessing the lexical complexity of short texts produced by second language learners of English.

P_Lex assumes that difficult words are infrequent occurrences not found in the 1,000 most common frequent words. It has a passing resemblance to the *Lexical Frequency Profile* (LFP), in the sense that words are sorted on the basis of a frequency list. However its authors claim that P_Lex seems to have some advantages over the LFP. Among other things, Meara and Bell (2001) claim that P_Lex works well with short texts, whereas the LFP requires texts over 200 tokens to obtain stable scores. However Meara and Bell advise to analyse texts whose minimum text length is 120 words, as results may not be stable below that minimum threshold. Bell (2003) points out that it does not mean that we cannot use P_Lex to analyse text shorter than 120 words, but it may undermine our degree of confidence in results.

> P_Lex divides the text into segments of ten words each, and then provides a profile showing the proportion of 10-word segments containing zero difficult words, the proportion containing one difficult word, two difficult words, so on and so forth, up to ten. When the text is processed, each word is compared

against the contents of the dictionary files. When words not included in any of the dictionary files are encountered by the computer (e.g. because they have inflectional or derivational affixes, or are not included in the frequency lists), the researcher is asked to allocate them to the correct band: mistake, name, number, Level 0 word, easy word (i.e. 1K word) or hard word (i.e. Beyond 1K) (see Figure 1).

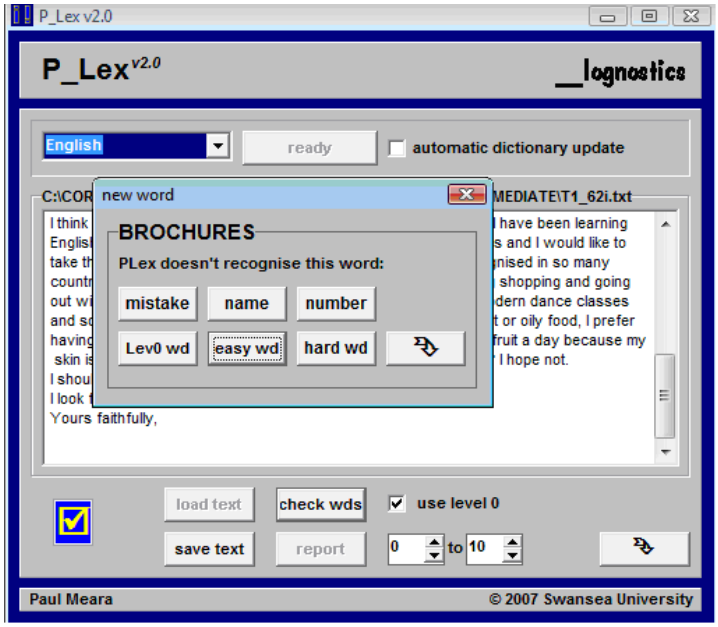

Figure 1 Example of a word not recognized by P_Lex

Before analysing the text, Bell (2003) points out that it should be decided whether *Level 0 words* which consist of 28 structure words, subsuming determiners, the most common pronouns, and past forms and participles of the verbs *do*, *have* and *be*– are to be included in the analysis or not. He advises to include them, since by doing that the amount of data to be analysed would be extended, a serious consideration when dealing with short texts. At the end of the analysis, *P_Lex* provides the following information (see Figure 2): (a) the number of tokens in the text; (b) the number of 10-word segments identified and processed; (c) the lambda value for the text, which is a single figure that indicates how likely the occurrence of difficult words is; and (d) an error value, which points to how close is the match between the lambda value displayed, and the Poisson distribution generated from lambda. The smaller the error value, the more satisfactory the match. Longer texts tend to produce smaller error values than shorter texts. Lambda values usually range from .5 to 4.5. It is assumed that the higher the lambda value, the bigger the vocabulary size of the learner (Meara & Bell 2001).

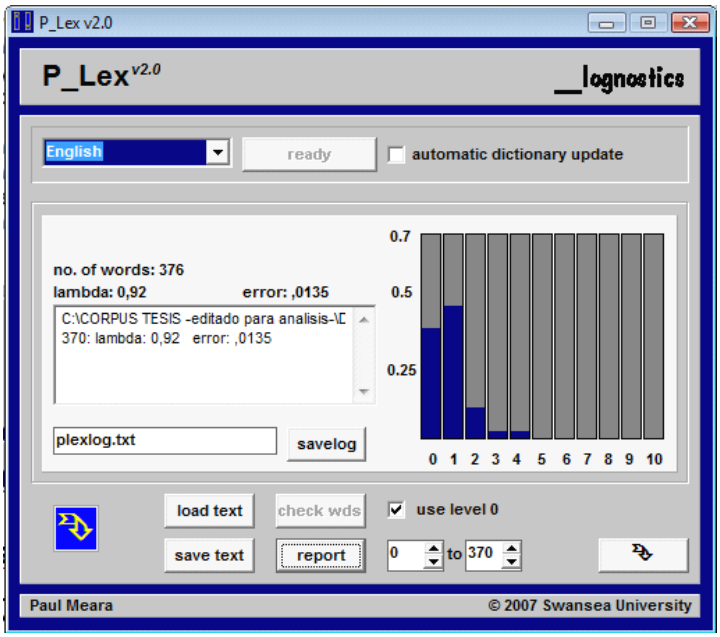

Figure 2 Example of a report produced by P_Lex

As Bell (2003) notes, it is advisable to pre-edit the texts before using P_Lex. As it does not identify multi-word items, he suggests identifying lexical phrases by connecting them with an underscore to classify them as difficult words, once the computer asks us to allocate them to the correct band. Furthermore, the editing process also includes correcting minor spelling mistakes. As Moreno *et al.* (2005) point out, 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. With regard to the different studies that have made use of this

electronic instrument, Moreno *et al.* (2005) note that some of them have been carried out with heterogeneous samples of informants from different proficiency levels and L1 backgrounds (e.g. Bell 2003; Meara & Bell 2001). This issue may prejudice results, as they are sensitive to learner variables (Farhardy 1982).

Investigations that have made use of P_Lex range from those that: (a) have analysed whether P_Lex is reliable across administrations (Bell 2003; Meara & Bell 2001); (b) have compared it against other lexical richness measures (e.g. Bell 2003; Daller & Xue 2007; Meara & Bell 2001; Miralpeix 2008; Read 2005); (c) have compared it against human raters (e.g. Daller & Phelan 2007; Moreno *et al.* 2005); (c) have explored whether it discriminates significantly between proficiency levels (e.g. Bell 2003; Meara & Bell 2001; Miralpeix & Celaya 2002; Moreno *et al.* 2005; Read & Nation 2006).

Thus, taking into account the previous research and its promise as an assessing tool that could work well with shorter texts, we consider it may be a suitable computer-mediated assessing tool to analyse learners' productive vocabulary in compositions.

V_Size

V_Size (Meara & Miralpeix 2007) is part of on-going research into ways of assessing the productive vocabulary of L2 learners. V_Size works on the assumption that "texts generated from a vocabulary of a particular size will tend to have a characteristic shape" (Meara & Miralpeix 2007, 1). Thus, it is assumed that that learners' text production can be modelled by weighting each word to its frequency and selecting words at random from a weighted list (Miralpeix 2008).

Hence, as Miralpeix (2008) points out that the programme generates sets of idealised lexical profiles, based on different vocabulary sizes (e.g. vocabularies of 1,000 words, 2,000 words, so on and so forth). Such profiles are generated by choosing words at random using the following logarithmic transformation of frequency [Ln(rankfreq)*1000]. Then V_Size finds the theoretical profile that best matches the lexical profile obtained from learners's texts – through a curve fitting approach – in order to provide an estimation of the vocabulary size of the test taker, on the basis of the text processed.

Thus, Meara and Miralpeix (2007) consider that the production of a learner can be matched against a model so as to infer the productive vocabulary size of the learner who produced that text. V_Size allows using different frequency lists as a yardstick criterion

to measure learners' vocabulary. Thus, it can be used: (a) an adapted version of the JACET list, which contains words deemed to be useful for learners; (b) the BNC list; or (c) researchers' own frequency dictionary built within the tool. Nation (2004) points out the BNC should not be used to assess secondary school learners, as the goals of the corpus from which the list has been compiled are different from the goals of the learners. Therefore, following Nation (2004), we believe that any frequency list compiled for educational purposes should be used by EFL teachers. It should be noted that V_Size is a recently developed experimental tool, which has already been used in the Spanish context by Miralpeix (2008) and Moreno and Jiménez (2008) with secondary school learners.

The authors of V_Size claim that their instrument goes beyond the profile reported by other tools such as Laufer and Nation's (1995) *Lexical Frequency Profile*, as not only does it provide a lexical frequency profile, but it also estimates what the profile tells us about the productive vocabulary size of the person who produced that text.

As aforementioned, V_Size allows choosing the frequency list to be used as the baseline of the analysis. The selection of the frequency list is an important decision to make, since as its authors note, it may affect the profile, giving as a result different shapes of the

curve. Thus, V_Size generates a profile for a specific text on the basis of five bands: Band A comprises the first 500 most frequent words in English; Band B includes the second 500 most frequent words in English; Band C and Band D comprise the third and fourth 500 most frequent words respectively, and the remaining words, which are considered to be infrequent words, are all categorised into Band E. Thus, when the text is processed, the programme generates a word list which includes all the words that are not recognised by the programme, so that the researcher and/or teacher can reclassify them manually (see Figure 3). It should be noted that the JACET list that V_Size uses has been partially lemmatised, which could pose some problems if the categorisation procedure is not done consistently.

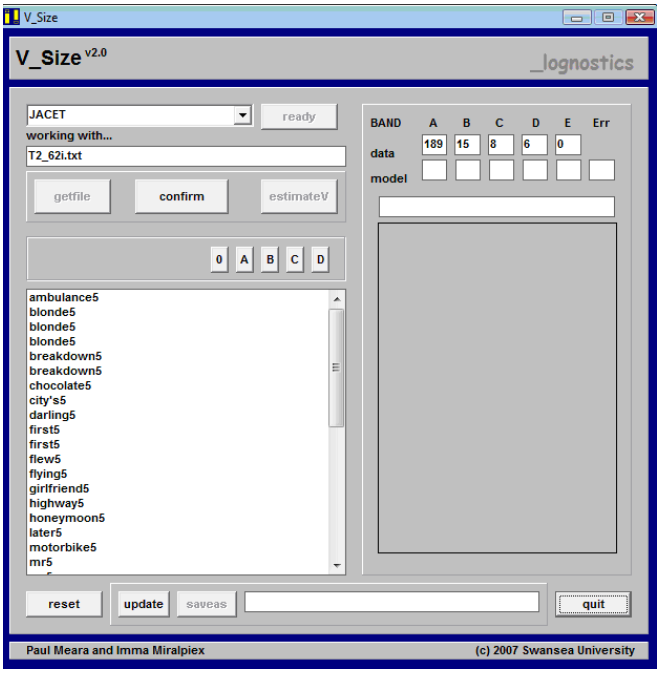

Figure 3 Example of words not recognised by V_Size

Meara and Miralpeix (2007) advise to reclassify proper nouns and numerals as Band A. Once the user confirms the reclassification of all words, the programme converts raw numbers into percentages. Afterwards, it compares the generated profile to a series of theoretical profiles stored in its memory and it reports which is the best match to the actual profile and estimates the vocabulary size of the author of the text (see Figure 4).

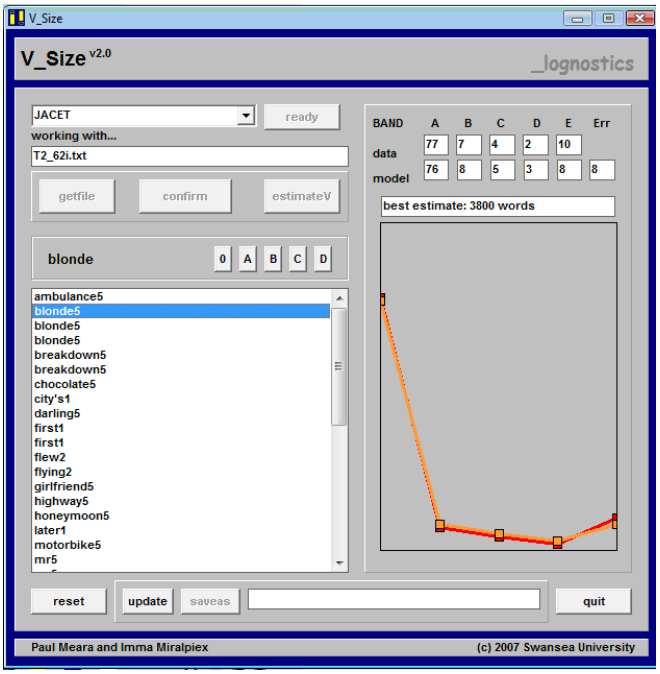

Figure 4 Example of the output produced by V_Size

From our point of view, despite V_Size is still and experimental tool, it offers the promising potential of estimating the productive vocabulary size of test takers, which may be very useful for EFL teachers to determine whether learners would be able to comply with communicative tasks successfully. Furthermore, it provides a further insight into L2 productive vocabulary use.

Conclusion and Pedagogical Implications

In this paper, we have addressed the importance of vocabulary for EFL learners, and we have reviewed a sample of electronic vocabulary analysers that EFL teachers may use to assess learners' vocabulary in compositions. Such computer-mediated tools offer complementary, rather than antagonistic views of vocabulary use in learners' compositions.

It is widely acknowledged that human raters are subjective, and neither a single composition per learner nor a single rater per composition produce reliable results to analyse the writing ability of test takers. For instance, scholars like Santos (1988) and McNamara (2000) point out that despite the use of rating schemes, and careful training, rating implies some kind of subjectivity. We should also mention the 'halo effect' which "is the effect of a feature which is not being tested, but which changes or influences the results" (Richards 1985, p. 128). Thus, for example, teachers that consider that a student with an immaculate behaviour in class will have a superb language performance are suffering from the halo effect, in other words, teachers' expectations may affect students' grades (Bachman 1990). Although subjectivity is an intractable problem of writing assessment, there are some measures that can reduce it (Bachman 2004; Henning 1987).

Hence, to obtain reliable results, each test taker should write two compositions and should be assessed by at least two raters (Jacobs *et al.* 1981).

The truth is that scoring essays in a reliable way is time-consuming and costly; and computers have played an important role for cost-reduction (Kaplan *et al*. 1998). Our view is that both human raters and automated raters should be seen as complementary rather than antagonistic entities. Should a human rater make use of an objective score elicited

Nowadays we come across different web-based commercial essay assessment systems for writing instruction (see Attali 2004; Burstein *et al.* 2003). For instance, *Criterion* has two applications: *e-rater*, which is an automated essay scorer; and another application called *Critique*, which comprises a suite of programmes that assess errors in grammar, usage, and mechanics, amongst other things (Burstein & Higgins 2005). Thus, as Burstein *et al*. (2003) point out, *Criterion* combines automated essay scoring and writing feedback, as an aid in writing instruction.

The problem is that using this kind of commercial systems in countries such as Malaysia or Spain requires a budget that high schools may not have. Hence, our goal has been to describe a sample of free electronic-computer mediated instruments, which could be used as a second rater to help EFL teachers in their assessment. In high-stakes examinations, essays present a practical problem and teachers have pressing schedules to rate compositions. If they could use any computer-mediated assessing instrument as a second rater, time and effort would be reduced.

Acknowledgements

This paper has been funded by Comunidad Autónoma de La Rioja and the Spanish Ministry of Science and Technology and FEDER (BFF2003-04009-C02-02, HUM2006-09775-C02-02). These grants are hereby gratefully acknowledged.

References

Agustín Llach, M. P. & Terrazas Gallego, M. (2009). Examining the relationship between receptive vocabulary size and written skills of primary school learners. *Atlantis, 31*(1), 129-147.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. Bristol: Continuum.

Astika, G. G. (1993). Analytical assessments of foreign students writing. *RELC Journal, 24*(1), 61-72.

Attali, Y. (2004). Exploring the feedback and revision features of *Criterion***.** Paper presented at the National Council on Measurement in Education (NCME) held between April 12 to 16, 2004, in San Diego, CA.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning, and Assessment, 4*(3), 3-29.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bauer, L. & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

Beglar, D. & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing, 16*, 131-162.

Bell, H. (2003). *Using frequency lists to assess L2 texts*. University of Wales Swansea: Unpublished PhD Thesis.

Bogaards, P. (2000). Testing vocabulary knowledge at a high level: The case of the Euralex French Tests. *Applied Linguistics, 21*(4), 490-516.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition, 23*, 321-343.

Burstein, J., Chodorow, M. & Leacock, C. (2003). *CriterionSM* online essay evaluation: An application for automated evaluation of student essays. Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence, held in Acapulco, Mexico, August 2003.

Burstein, J. & Higgins, D. (2005). Advanced capabilities for evaluating student writing: detecting off-topic essays without topic-specific training. 12th International Conference on Artificial Intelligence in Education, held July 2005 in Amsterdam, the Netherlands

Carter, R. (1998): *Vocabulary*. London: Routledge.

Cobb, T. (2000). One Size Fits All? Francophone learners and English vocabulary tests. *The Canadian Modern Language Review, 57*(2), 295-322.

Coniam, D. (1999). An investigation into the use of word frequency lists in computing vocabulary profiles. *Hong Kong Journal of Applied Linguistics*, *4*(1), 103-23.

Daller, H. & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 234-244). Cambridge: Cambridge University Press.

Daller, H. & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, and J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 150-164). Cambridge: Cambridge University Press.

Daller, H., Milton, J. & Treffers-Daller, J. (2007) (Eds.). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

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

Farhady, H. (1982). Measures of language proficiency from the learner's perspective. *TESOL Quarterly*, 16 (1), 43-59.

Goethals, M. (2005). WordClassifer. Version 2.5. *TESL-EJ, 9*(1), 1-7.

Henning, G. (1987). *A Guide to Language Testing. Development. Evaluation. Research*. Massachusetts: Newbury House Publishers.

Hirsch, D. & Nation, I.S.P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure?. *Reading in a foreign language, 8*(2), 689-696.

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

Jiménez Catalán, R. M. & Moreno Espinosa, S. (2005): Promoting English vocabulary research in primary and secondary education: Test review and test selection criteria. *English Studies International Peer-Reviewed Scholarly Journal, 26*, 171-188.

Jiménez Catalán, R. M. & Mancebo Francisco, R. (2008). Vocabulary input in EFL textbooks. *Revista Española de Lingüística Aplicada (RESLA), 21*, 147-165.

Kaplan, R.M, Wolff, S.E., Burstein, J. C., Lu, C., Rock, D. A. & Kaplan, B. A. (1998). Scoring essays automatically using surface features. ETS Research Report 98-39. Princeton, N J: Educational Testing Service. (GRE Report No. 94-21P).

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126-132). London. MacMillan.

Laufer, B. (1997). The lexical plight in second language reading. In J. Coady, & T. Huckin, (Eds.), *Second language vocabulary acquisition* (pp. 20-34). Cambridge: Cambridge University Press.

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

Laufer, B. (2001). Quantitative evaluation of vocabulary: How it can be done and what it is good for. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, & T. McNamara (Eds.), *Studies in language testing 11: Experimenting with uncertainty* (pp. 241-250). Cambridge: Cambridge University Press.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.

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

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

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Meara, P. (1980). Vocabulary Acquisition: A neglected aspect of language learning. *Language Teaching and Linguistics, 13*(4), 221-246.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams, (Eds.), *Performance and competence in second language acquisition.* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26*(1), 32-47.

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

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

Meara, P. & Miralpeix, I. (2007). *V_Size*. University of Wales Swansea.

Miralpeix, I. (2008). *The Influence of Age on Vocabulary Acquisition in English as a Foreign Language*. PhD Thesis. University of Barcelona.

Miralpeix, I. & 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.), *Proceedings of the XXVI AEDEAN Conference*, (pp. 399-406). Santiago de Compostela: Santiago de Compostela University Press.

Moreno Espinosa, S., Fernández Fontecha, A. & Agustín Llach, M. P. (2005). Can P_lex Accurately Measure Lexical Richness in the Written Production of Young

Learners of EFL?. *Porta Linguarum: Revista Internacional de Didáctica de las Lenguas Extranjeras, 4*, 7-21.

Moreno Espinosa, S. & Jiménez Catalán, R. M. (2008) Measuring the productive vocabulary of EFL learners at the beginning of secondary education. Paper presented at the *41st British Association of Applied Linguistics Conference* (BAAL 2008). Swansea University, United Kingdom.

Muncie, J. (2002). Process writing and vocabulary development: Comparing lexical frequency profiles across drafts. *System, 30*, 225-235.

Nagy, W. & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, *19*, 304-330.

Nation, I. S. P. & Coady, J. (1988). Vocabulary and reading. In R. Carter & M. McCarthy (Eds.), Vocabulary and language teaching (pp. 97-110). London: Longman.

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

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I.S.P. (2004). A study of the most frequent word families in the British National Corpus. In B. Laufer & P. Bogaards (Eds.), *Vocabulary in a Second Language* (pp. 3-13). Amsterdam: John Benjamins Publishing Co.

Nation, I.S.P. (2005). *Instruction Manual for Range and Frequency*. Available from http://www.vuw.ac.nz/lals/

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening*? The Canadian Modern Language Review, 63*(1), 59-81.

Nation, I.S.P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 35-43). Cambridge: Cambridge University Press.

Nation, I.S.P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Nattinger, J. R. & DeCarrico, J. S. (1997). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles M. E., Kukich, K. (2002). Stumping *e-rater*: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*, 103-134.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19*(2), 12-25.

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

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined?. In P. Bogaards & B. Laufer (Eds.): *Vocabulary in a Second Language* (pp. 209-227). Amsterdam: John Benjamins.

Read, J. (2005). Applying lexical statistics to the IELTS listening test. *Research Notes, 16*, 12-16.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1-32.

Read, J. & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. In P. McGovern & S. Walsh (Eds.): *IELTS research reports* 6 (pp. 207-231). Canberra: IELTS Australia.

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*(1), 77-89.

Richards, J. (1985): *Longman dictionary of Applied Linguistics*. Harlow: Longman.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*, 69–90.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*(2), 281-317.

Schmitt, N. and Meara, P. (1997). Researching vocabulary through a word knowledge framework — word associations and verbal suffixes. *Studies in Second Language Acquisition, 19*, 17–36.

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

Schmitt, N. & Zimmerman, C. B. (2002). Derivative word forms: What do learners know?. *TESOL Quarterly, 36*(2), 145-171.

Taylor, C., Kirsch, I. & Eignor, D. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning, 49*(2), 219-274.

Vermeer, A (1992). Exploring the second language learner lexicon. In L. Verhoeven & J. H. A. L. de Jong (Eds.), *The construct of language proficiency*. *Applications of*

psychological models to language a*ssessment* (pp. 147-162). Amsterdam: John Benjamins.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 173-189). Amsterdam: John Benjamins Publishing Company.

Webb, S. & Rodgers, M.P.H. (2009). The Lexical Coverage of Movies. *Applied Linguistics, 30*(3), 407-427.