ARTÍCULO DE INVESTIGACIÓN • VÍNCULOS-ESPE (2023) VOL.8, No.1:41-60
DOI: https://doi.org/10.24133/vinculosespe.v8i1.3114

REVISTA VÍNCULOS
UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

Analysing Vocabulary in Human Graded L2 Scripts Through Automatic Lexical Analysers

Análisis del vocabulario de alfabetos L2. A través de analizadores léxicos automáticos

JOSÉ LEMA ALARCÓN

Universidad de las Fuerzas Armadas – ESPE
Av. General Rumiñahui s/n y Ambato, Sangolquí – Ecuador

jslema@espe.edu.ec

ABSTRACT

In grading second language (L2) scripts, teachers take approximate measures regarding lexical choices that may suggest the overall quality of the texts. This study compared various measures of lexical proficiency in scripts written by English language students. The corpus was analyzed to determine the correlation between teacher judgments and lexical items in L2 written assignments. Using the Text Inspector online tool, the study first attempted to delimit the lexical items corresponding to levels (e.g., A1/2, B1/2, and C1/2) of the Common European Framework of Reference for Languages (CEFR). In addition to verifying vocabulary levels, the study also used the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) to analyze advanced words and phrases used in each script. Using the Text Inspector tool, the first part of the study demonstrated that the assigned grades for each script correlated with the CEFR word lists. Similarly, the grades and L2 scripts were correlated with twenty-two indices of lexical sophistication (i.e., academic word frequency, range and N-gram proportion frequency).

Keywords: Lexical sophistication, CEFR vocabulary levels, Text Inspector, TAALES, English Vocabulary Profile EVP, vocabulary, writing proficiency, grading.

Recibido: 2022-10-15
Aceptado: 2022-12-15

INTRODUCTION

As an additional dimension to lexical competence, lexical sophistication comprises the accessible breadth and depth of the lexical knowledge of L2 writers. A variety of frequency measures, which are compared with representative corpora (e.g., Cambridge Learner Corpus), have been developed to measure the size and quality of the lexical inventories of English language students. In addition to individual word frequency counts, frequency measures of word associations (e.g., n-grams), academic lists (e.g., words and formulas), and psycholinguistic properties (e.g., Familiarity and Age of Acquisition) have arisen as estimates of L2 student's lexical sophistication (Laufer & Nation, 1995; Meara, 1996a; Simpson-Vlach & Ellis, 2010; Crossley, Cai, & McNamara, 2012; Kyle & Crossley, 2015). The current study attempts to measure vocabulary proficiency levels as reflections of the quality and judgment of L2 written production (Crossley & McNamara, 2011; Nation, 2001; Olinghouse & Wilson, 2013). In judging L2 written assignments, both novice and experienced teachers may find it difficult to determine whether a word or phrase is basic, intermediate, or advanced and whether the use of individual words and phrases can lead to lexical sophistication and writing proficiency. Acknowledging this, the English Vocabulary Profile (EVP) project (Capel, 2010, 2011, 2012; Saville, 2012; Saville & Hawkey 2010) asserts that lexical variety in scripts can be measured and aligned to CEFR vocabulary levels (Lenko-Szymanska, 2015). In addition to verifying vocabulary in scripts, Attali and Burstein (2006) discuss the use of n-grams, word frequencies, and range to help the teacher gauge lexical ability. More recently Kyle and Crossley (2015) suggested that writing and lexical proficiency can be predicted using fifty-five indices of lexical sophistication. Having said that, the current study attempts to clarify the association between the grades assigned by teachers and vocabulary choices in formative written assignments or scripts by young adult English as a Foreign Language (EFL) students.

Knowing a Word

In a seminal article on the role of vocabulary teaching, Richards (1976) posed the still unanswered question, "What does it mean to know a word?" Despite the fact that Richard's article reflected the vocabulary research of the time, he proposed a conceptual framework to find answers relevant to the teaching of vocabulary. According to this framework, understanding a word was associated with knowing its frequency, register, syntactic behavior, form and derivations, network of associations, semantic value, and word senses. Over the years, the framework has been expanded and scrutinized by other scholars interested in the issue of lexical knowledge (Nation 1990; Schmitt, 1995, 1998; Schmitt & Meara, 1997). Yet, Meara (1996b) noted that the framework proposed by Richards (1976) for 'knowing a word' is an impossible one and that applying frameworks that attempt to characterise individual words is futile. Instead, L2 vocabulary should be looked at through the properties of lexical units as a whole. Admittedly, the ideas described above continue to be influential and inspirational in the quest to explain the nature of vocabulary teaching, however, new technological developments (e.g., Natural Language Processing) have enabled researchers to investigate lexical aspects that would have been impossible two decades ago. The next section discusses the development of language levels set by the CEFR, the English Profile project, and technological advancements in the EVP.

REVISTA VÍNCULOS UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

English language levels and the Common European Framework Levels

As in other contexts where English is learned and taught as a foreign language, Ecuador has been strongly influenced by CEFR in the teaching, learning, and evaluation of languages. Indeed, the CEFR levels and standards, aimed to describe language skills based on criteria set by the Council of Europe (Council of Europe, 2011), have permeated all aspects of the teaching and learning of English. For example, the CEFR standards are inherent to the syllabus design (i.e., course distribution aligned with A1-C2 standardized levels), and the assessment of language attainment such as the Cambridge English (i.e., Key English Test (KET), Preliminary English Test (PET), and First Certificate in English (FCE) testing formats (British Council, 2015). Incidentally, we need to recognize that the CEFR (2011) presents not only proficiency scales and levels through Reference Level Descriptors (RLD) for languages, but also describes to some extent what users of a language can do in various communicative activities and tasks. One of the most important RLD developments has been the Threshold Level (van Ek, 1980) intended to explain what L2 learners can do and what functions they can perform with the English language. More specifically, the English Profile project has attempted to provide details on what learners should be able to do with English vocabulary at A1- B2 and C1-C2 levels (Capel, 2010, 2012).

The English Vocabulary Profile

English Profile project efforts to further the implementation of RLDs have led to the development of vocabulary lists (by level, category, and 'can do' statements) that are now included in the EVP and English Grammar Profile projects (Capel, 2010, 2012; Carter, McCarthy, Mark, & O'Keeffe, 2011). For example, the EVP interactive online resource allows teachers, students, and researchers to determine which lexical items and phrases are typically mastered by learners at each CEFR level (Capel, 2010; Harrison & Barker, 2015). Additionally, Amkham (2016) used Text Inspector (Bax, 2015) to verify selected vocabulary of L2 students who were given the EVP as a supporting tool during their writing process. Even though Amkham's study found no statistical differences in the number of B2, C1, and C2 words combined in tasks written with and without the help of the EVP, they did find some increase in the number of sophisticated words in students' writings at the C1 and C2 levels when the texts were analyzed separately. Negishi, Tono, and Fujita (2012) analyzed the difficulty factors of phrasal verbs included in the EVP and concluded that some lexical items described in the EVP are either inaccurately ordered or the vocabulary items had overlapping meaning. In addition, the EVP and Text Inspector have been used to analyze glossaries in textbooks and to check whether their vocabulary aligns with CEFR levels (Millar, 2016). In the first part of this paper, I analysed L2 scripts using the online Text Inspector software. In the process, I considered previous studies in which the EVP was used to analyze and compare student's vocabulary development and validate the proper use of EVP vocabulary levels. Furthermore, in the second part of the study, L2 scripts were examined using various indices of lexical sophistication as presented by the Tool for the Automatic Analysis of Lexical Sophistication (TAALES, Version 2.2; Kyle & Crossley 2015).

Indices of lexical sophistication

Word frequency

The frequency of a word can be understood as follows: the more recurrent the language (words and phrases) contained in a corpus, the easier the words will be to learn and the earlier they will be used (Schmitt, 2012, p. 71). For example, in academic writing, words need to be carefully selected from the low-frequency writer's repertoire, while in ordinary speech high-frequency or more ordinary words might be the norm. In addition to this, words that contain more syllables increase processing time (Balota, Cortese, Sergent- Marshall, Spieler, & Yap, 2004). The implication here is that high-frequency words carrying fewer syllables are retrieved faster than low-frequency words and this selection of a less frequent, albeit more complex word, indicates a better predictor of lexical and writing ability (Kuperman, Stadthagen-Gonzales, & Brysbaert, 2012).

Word range

Measuring word range can be another predictor of lexical sophistication and writing quality. Word range can be contrasted with frequency in the sense that range or dispersion refers to how widely the word occurs across different texts within a corpus. In addition, it is suggested that if a lexical unit is regularly more dispersed and occurs across many texts, the word is likely to be a high-frequency one (Gries, 2008; Kyle & Crossley, 2015. p. 760).

N-grams, frequency and range

One such area that has not been thoroughly investigated in L2 writing is the production of multi-word units (e.g., n-grams). Such units are of interest because they provide both lexical and syntactic information about a text. N-grams refer to groups of two or more words that repeatedly appear in language as fixed items more frequently than expected by chance and much more frequently than phrasal verbs and idioms. In addition, n-gram frequency and range share the same logic of word range and frequency stated above, that is less frequent and dispersed n-grams are better predictors of lexical sophistication and writing proficiency (Hyland 2008; Kuperman et al., 2012; Crossley et al., 2012). Stubbs (2007) defines an n-gram as a 'recurrent string of orthographic word forms' (p. 166), while for Chen and Baker (2014), it is 'continuous word sequences retrieved by taking a corpus-driven approach with specified frequency and distribution criteria.'

Bigram and trigram association strength

Gries and Ellis (2015) contend that from a psycholinguistic point of view, language learning is 'associative learning of representations that reflect the probabilities of occurrence of form-function mappings' (p. 4). The association between words is measured to determine the extent of a recurrent string of orthographic word forms. Kyle and Crossley (2015) include various associations measuring for n-gram indices such as, mutual information for low-frequency words (e.g., 'by degrees') and t-score for high-frequency words (e.g., 'nowhere to'). In addition to measuring

REVISTA VÍNCULOS UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

n-grams, Delta P attempts to elucidate the asymmetry between words, for example the lexeme Achilles predicts heel or tendon more strongly to the right, rather than heel or tendon predicts Achilles to the left, thus the association between Achilles and heel or tendon is very strong, but Achilles tendon (F: 202) is more frequently used than Achilles heel (F:163) or Achilles area (F:2). The Delta P measure is relevant because it helps to predict different n-gram associations (Allan, 1980; Gries, 2013); lastly, the approximate collexeme index uses mutual information, and Delta P measures to better clarify word combinations and higher lexical proficiency (Durrant & Schmitt, 2009).

Bigram and trigram proportions

Kyle and Crossley (2015) also developed indices that depend on the co-occurrence of words in a corpus. They suggest that L2 proficient students tend to use a higher proportion of bigrams and trigrams found in the Corpus of Contemporary American (COCA) n-gram lists (Davies, 2008). For example, scripts that include bigrams and trigrams in the 10k to100k most frequent n-grams of the COCA are more likely to have better human judgments or higher grades. The n-gram indices are intended to report on the proportion of bigram and trigram scores in scripts and in the corpus. Crossley et al., (2012) argue that higher rated essays contain proportionally fewer bigrams and trigrams 'in much the same way that higher quality essays contain a smaller proportion of frequent words' (p. 216).

Current study

Building on from the idea that knowing a word entails countless dimensions, many efforts have been made to explain words separately and collectively; those attempts include measures such as frequency and range counts, n-gram association measures, and frequency proportion counts. Thus, this study attempts to clarify the relationship between teachers' judgments and the vocabulary used in scripts produced by EFL students.

Because the level of the scripts collected for this study is mainly aligned with the vocabulary and 'can do' statements set by CEFR (2011) benchmarks, the author first considers whether the scripts meet these standards and relate this with the grades assigned by the teachers. In addition, to measure and determine the CEFR levels in each script, the Text Inspector analysis tool and the EVP are used (Capel, 2010; Saville, 2012; Saville & Hawkey, 2010). Once the CEFR vocabulary and writing levels were verified, the lexical items in each script were measured utilizing the TAALES and fifty-five indices for measuring lexical sophistication in writing (Kyle & Crossley, 2015).

Purpose statement and research questions

To understand the role of lexical sophistication and writing proficiency, this quantitative study aims to provide a discussion of the lexical items included in language learners' written assignments and whether they correlate with the grades assigned by their teachers. The work also attempts to analyze the relationship between variables that may influence the quality of written texts. Thus, the study addresses two specific research questions.

1. What is the relationship between written-graded scripts and lexical assessment criteria?

2. When using lexical sophistication tools to measure the vocabulary within scripts, what is the relationship between the grades and the indices of lexical sophistication?

Answering these questions might enable the researcher not only to better understand the elements deemed by the teacher when assigning a grade in L2 writings, but finding answers to those research questions can also help language teachers to raise awareness about the different dimensions in the teaching, learning and assessing of vocabulary. For example, in the development of practical teaching strategies, teachers may include explanations, and develop pedagogical strategies emphasizing that vocabulary comprises different lexical levels of sophistication; that not only frequency and range of single words play a key role in vocabulary selection, but also the combination of multiword or phrases (e.g., bigrams and trigrams); that the context needs consideration during the selection of particular lexemes or lexical units; and pointing out the fact that when selecting specific words and phrases (e.g., academic vocabulary) in their writings, the level of vocabulary sophistication, might affect the overall quality of a text.

METHODOLOGY

Research background

The scripts were written by undergraduate L2 students enrolled in a variety of academic programmes (i.e., engineering, law, medicine, and teaching). Data gathering for this study involved formative, graded, written texts aimed to support the students' language learning development. The three institutions where the data were collected utilise online Virtual Learning Environments (VLE), such as Moodle, to facilitate delivery and provide feedback and grading of written assignments.

The collected scripts were typed using the institutional VLE platform and consisted of formative writing tasks and writing topics such as personal opinion compositions, emails, letters of application, letters of complaint, different recommendations, short stories, curriculum vitae, book and film reviews, reports, blog entries. Table 1 below details the number of scripts collected in each institution, the number of courses, the total number of writers per institution, the number of texts, and the number of topics written. All of the texts and topics collected are aligned and follow the criteria set by the CEFR B1 level.

Table 1
Collected scripts: number of institutions, courses, texts, writers and topics

	Courses	Writers	Texts	Topics
Institution 1	2	40	10	10
Institution 2	1	11	30	6
Institution 3	6	120	60	2

Grades

Each script was reviewed and graded by qualified teachers with significant EFL experience, which adds an extra element of validity and reliability to the study. The grades were the de-

REVISTA VÍNCULOS UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

pendent variable used to determine the relationship between the grades assigned to each script and the subscale range of language—vocabulary. This is important because teachers had already considered the assessment aspects of each script (e.g., learning objectives, outcomes, scales, and subscales). Thus, the present study focused its analysis on verifying students' vocabulary at the CEFR levels and on indices of lexical sophistication for writing quality. Consequently, as shown in Table 2 below, the mean of all grades is greater than seven points, suggesting that most scripts exceeded the 70 percent required in the three institutions for any assignment to be considered satisfactory.

Table 2
Descriptive statistics, 35-word threshold samples: mean, standard deviation

Corpus	N	Grade mean	SD	Word count mean	SD
Group 1	208	8.17	1.10	153.04	61.11
Group 2	228	8.02	1.04	66.84	29.26
Group 3	200	8.90	1.26	172.18	70.52

Index measures

Text Inspector, English Vocabulary Profile, and CEFR levels

To verify the vocabulary levels of each script, the online text analyzer, Text Inspector, and the Cambridge English Vocabulary Profile EVP function tool were used. (Saville, 2012; Saville & Hawkey, 2010). Even though the Text Inspector online tool offers several types of analyses (e.g., lexical diversity, tagger, spelling errors, BNC/COCA/AWL lexis, metadiscourse, and scorecards) this study focused on the Cambridge EVP. The EVP wordlist was developed as part of the English Project and enables users to verify the difficulty of words in a text. The EVP function of the Text Inspector specifies the words of a text at each CEFR level. The Lexis EVP function bases its analysis on word frequency and word sense and has been informed and supported by corpora sources such as the Cambridge English and Cambridge Learner corpora (Nicholls, 2003; Granger, 2008).

Text Inspector categorises each word by level according to its meaning and the sense given to the context of the word. Thus, a word can have a different meaning in different contexts and might belong to more than one different category or level. For example, the word bank is an A1-level word when referring to a financial institution and a B2-level word when referring to an area of land adjacent to a body of water. In the text analysis, however, the software recognizes only the A1 sense of the word. Text Inspector, as noted above, selects the lowest-level meaning of a word by default. This is important to consider when analysing many texts because many words and scripts will be analysed as if they are using entirely low-level or high-frequency words.

Analyzing word types with Text Inspector and Lexis EVP

Word type verification (e.g., function and content words) in L2 students' writings and their corresponding alignment with CEFR levels was done using Text Inspector and the Lexis EVP function. The analyses were done on three groups of scripts each. The results of the analysis are shown in Table 3.

Table 3
Descriptive statistics of word type counts using Text Inspector and Lexis: CEFR Levels, mean, and standard deviation.

		A1	A2	B1	B2	C1	C2	Total
Group 1	Mean	50.4	11.7	8.9	4.0	0.8	0.4	76.3
	SD	15.4	6.0	5.4	3.5	1.1	0.9	24.9
	%	**66%**	**15%**	**12%**	**5%**	**1%**	**1%**	**100%**
Group 2	Mean	29.4	6.0	3.5	1.6	0.2	0.2	40.9
	SD	9.4	3.3	2.6	1.5	0.6	0.4	13.5
	%	**72%**	**15%**	**9%**	**4%**	**-**	**-**	**100%**
Group 3	Mean	52.2	12.8	12.2	6.4	1.7	0.5	85.8
	SD	17.6	6.5	8.3	5.3	2.0	0.8	32.9
	%	**61%**	**15%**	**14%**	**7%**	**2%**	**1%**	**100%**

- The results show that Group 3 had the greatest proportion of word types at B1, B2, and C1 levels (14, 7, and 2 percent, respectively), the lowest proportion at the A1 level (61). The proportion of word types at the A2 level (15) in Group 3 was similar to Groups 1 and 2.

- Compared to Group 3, Group 1 had fewer word types at the B1, B2, and C1 levels (12, 15, and 1 percent, respectively) and a higher number of word types at A1 (66).

- Group 2 had the lowest percentage, mean, and standard deviations of the three groups. Group 2 also had the most word types at the A1 level (72) and the least at B1, B2, C1, and C2 (9, 4, 0, and 0, respectively).

It is important to highlight the fact that the results are only estimates because Text Inspector does not include unknown words that might have been misspelt, mistyped or not included in the EVP word lists. The word type results (Table 3) also suggest that script length in each group might play a role in the number of word types per level. Typically, the more words a script contains, the greater the number of word types.

TAALES and indices of lexical proficiency

Even though Text Inspector counts the frequency of word types at all CEFR levels, the CEFR does not provide detailed information on what vocabulary indices or micro- features should be considered as predictors for advanced vocabulary and writing proficiency. The TAALES measures micro-features or indices that enable the prediction of lexical sophistication and academic writing quality, such as frequency and range, n-grams, and word proportions in academic scripts (Kyle & Crossley, 2015).

REVISTA VÍNCULOS
UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

Micro-features

Table 4
Groups and categories of indices analysed in TAALES for academic writing

Word Frequency and Range		
All Words	Content Words	Frequency Words
1. Academic Range AW	5. Academic Range CW	9. Academic Range FW
2. Academic Frequency AW	6. Academic Frequency CW	10. Academic Frequency FW
3. Academic Range Log AW	7. Academic Range Log CW	11. Academic Range Log FW
4. Academic Frequency Log AW	8. Academic Frequency Log CW	12. Academic Frequency Log FW
Bigram Frequency, Range and Association of Strength		
Bigram Frequency and Range	Bigram Association Strength	Bigram Proportion Frequency
13. Academic Bigram Frequency	17. Academic bi MI	22. Academic bi prop 10,000
14. Academic Bigram Range	18. Academic bi M2	23. Academic bi prop 20,000
15. Academic Bigram Frequency Log	19. Academic bi T	24. Academic bi prop 30,000
16. Academic Bigram Range Log	20. Academic bi DP	25. Academic bi prop 40,000
	21. Academic bi AC	26. Academic bi prop 50,000
		27. Academic bi prop 60,000
		28. Academic bi prop 70,000
		29. Academic bi prop 80,000
		30. Academic bi prop 90,000
		31. Academic bi prop 100,000
Trigram Frequency, Range and Association of Strength		
Trigram Frequency and Range	Trigram Association of Strength	Trigram Proportion Frequency
32. Academic Trigram Frequency	36. Academic tri MI	46. Academic tri prop 10,000
33. Academic Trigram Range	37. Academic tri MI2	47. Academic tri prop 20,000
34. Academic Trigram Frequency Log	38. Academic tri T	48. Academic tri prop 30,000
35. Academic Trigram Range Log	39. Academic tri DP	49. Academic tri prop 40,000
	40. Academic tri AC	50. Academic tri prop 50,000
	41. Academic tri 2 MI	51. Academic tri prop 60,000
	42. Academic tri 2 MI2	52. Academic tri prop 70,000
	43. Academic tri 2 T	53. Academic tri prop 80,000
	44. Academic tri 2 DP	54. Academic tri prop 90,000
	45. Academic tri 2 AC	55. Academic tri prop 100,000

The 55 micro-features used to predict lexical sophistication were taken from the list of indices provided by TAALES (Kyle & Crossley, 2015, p. 758). All of the micro-features selected for the study belong to the academic writing category, which is based and informed by the COCA database (Davies, 2008). The TAALES micro-features are described above.

Variables

- The scores provided by Text Inspector (CEFR levels) and those by TAALES (indices of lexical sophistication) were the independent variables (predictors).

- The criterion variable, or dependent variable, was the grade (outcome) each script was given by the teacher (1=low to 10=high). The study uses the same criterion variable in both of the correlation analyses described below.

Procedure

Data collection

Access was granted to three different Moodle institutional platforms where the grades and scripts were stored. Using a private, password-protected, computer, the data were downloaded to a file (that was later encrypted) and grouped to help further analysis and management. The data consisted of grades, scripts, and the location where the data were collected (see Table 2 above). In addition, each script was analysed using Text Inspector and TAALES. The statistical correlations of variables for the two sets of scores were divided in Studies 1 and 2.

RESULTS

Study 1: Results and discussion

Study 1 Analysis

The first set of scores failed the test of normality and assumes that the independent variable set is non-parametric. Cohen and Manion (1985) describe a variety of correlational methods for relationship measuring (such as rank order or Spearman's rho) that are designed to assist the researcher in finding the association of non-linear sets of scores that comprise ordinal scales. Spearman's rho correlation analysis allows us to associate non-parametric sets of data. Table 5 presents the correlation of a set of variables using Spearman's rho used to find the ordinal relationship between two sets of scores—the grades and vocabulary counts at each CEFR level.

Table 5
Spearman's rho correlation between grades and words at each level of the CEFR

	A1	A2	B1	B2	C1	C2
Correlation Coefficient	.105**	.045	.147**	.158**	.127**	.082*
Sig. (2-tailed)	.008	.260	.000	.000	.001	.038

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

REVISTA VÍNCULOS UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

A series of Spearman correlations were done on Study 1 data and five significant correlations were found. All five of these correlations were positive and weak, with the variables significantly associated with grades consisting of the EVP A1 type count, ρ(636)

=.105, p <.01; B1 type count, ρ(636) =.147, p <.01; EVP B2 type count, ρ(636) =.158, p

<.01; EVP C1 type count, ρ(636) =.127, p <.01; EVP C2 type count, ρ(636) =.082, p <.05. In all cases, higher values on these five measures were associated with higher grades.

Discussion Study 1

Initially and employing Text Inspector scores, a Spearman's rho (ρ) correlation coefficient study was carried out to determine the relationship of ordinal variables: grades and CEFR vocabulary levels. The grades and the types of words included in the scripts correlated at the A1, B1, B2, C1, and C2 levels. The correlations were found to be positive and weak and significantly associated with the grades.

Study 2: Results and discussion

Analysis Study 2

The original data included scripts ranging from less than 100 words to more than 200 words (See Table 2).

Table 6
Descriptive statistics, 100-word threshold sample: mean, standard deviation (SD)

Corpus	N	Grade mean	(SD)	Word count mean	(SD)
Group 1	172	8.16	1.08	168.68	56.30
Group 2	24	7.56	0.87	124.75	32.73
Group 3	185	8.94	1.47	178.75	68.65

Table 7.
Correlation between grades and fifty-five indices of lexical sophistication

Indices:COCA Academic Word Frequency and Range

All Words (AW)	rs	rs2	Content Words (CW)	rs	rs2	Function Words (FW)	rs	rs2
Range_AW	.108*	0.0117	Range_CW	.163**	0.0266	Range_FW	.171**	0.0292
Range_Log_AW	.157**	0.0246	Range_Log_CW	.201**	0.0404	Range_Log_FW	.193**	0.0372
Frequency_AW	.025		Frequency_CW	.025		Frequency_FW	.059	
Frequency_Log_AW	.142**	0.0202	Frequency_Log_CW	.217**	0.0471	Frequency_Log_FW	.123*	0.0151

Indices: COCA Academic Bigram Frequency, Range and Association of Strength

Bigram Frequency Range	rs	Association Strength	rs	rs2	Proportion Frequency	rs	rs2
Bigram_Frequency	.030	bi_MI	.022		bi_prop_10k	.133**	0.0177
Bigram_Frequency_Log	-.041	bi_MI2	-.008		bi_prop_20k	.132**	0.0174
Bigram_Range	-.037	bi_T	.036		bi_prop_30k	.133**	0.0177
Bigram_Range_Log	-.036	bi_DP	.137**	0.0188	bi_prop_40k	.156**	0.0243
		bi_AC	.034		bi_prop_50k	.147**	0.0216
					bi_prop_60k	.159**	0.0253
					bi_prop_70k	.160**	0.0256
					bi_prop_80k	.152**	0.0231
				–	bi_prop_90k	.159**	0.0253
					bi_prop_100k	.159**	0.0253

Indices: COCA Academic Trigram Frequency, Range and Association of Strength

Frequency and Range	rs	rs2	Association of Strength	rs	rs2	Proportion Frequency	rs	rs2
Trig_Fr	.019		tri_MI	.017		tri_prop_10k	.060	
Trigram_Range	.010		tri_MI2	.016		tri_prop_20k	.059	
Trigram_Frequency_Log	.002		tri_T	.020		tri_prop_30k	.071	
Trigram_Range_Log	-.011		tri_DP	.090		tri_prop_40k	.102*	0.0104
			tri_AC	.029		tri_prop_50k	.087	
			tri_2_MI	.098		tri_prop_60k	.086	
			tri_2_MI2	.078		tri_prop_70k	.090	
			tri_2_T	.031		tri_prop_80k	.087	
			tri_2_DP	.091		tri_prop_90k	.094	
			tri_2_AC	.018		tri_Prop_100k	.085	

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

However, Crossley and McNamara (2013) noted that analyses of scripts with few words produced unreliable results. In addition, there is the claim that word length is strongly correlated with essay quality and consistent analysis relies on scripts containing enough linguistic representation, that is, lexical, syntactic, and cohesion elements (Ferrand, Brysbaert, Keuleers, New, Bonin, Méot, & Pallier, 2011). Consequently, and to strengthen the reliability of Study 2, only texts of 100 words or more were included, and leaving a corpus of 381 samples (see Table 6 above).

Regarding these results, significant correlations were found between the grades and the three groups of scores obtained using TAALES (see Table 7 above).

Discussion of Study 2

Word frequency and range. The scores obtained using TAALES contained indices of lexical sophistication (i.e., frequency and range for content words, function words, and all words). The results show that frequency indices for all words, function and content words correlated positively with teachers' judgments. These results may imply that the scripts comprised high-frequency content and function words deemed as more frequent and therefore less sophisticated. Likewise, these outcomes appear unexpected because of the notion that negative correlations may suggest that higher-graded scripts contain less- frequent and more sophisticated words. In addition, these findings can be related to previous studies, for example Kyle & Crossley (2015) found positive correlations between holistic scores and frequency of content words in L2 learners who produced more frequent words after spending more time learning the language. They also suggest that the positive correlations are divergent from other study findings (Crossley, Salsbury, McNamara, & Jarvis, 2011a, 2011b) where negative correlations related to more advanced L2 writers producing low-frequency and more sophisticated words.

The results *for raw and log range* (e.g., content and function words) indices correlated positively with higher grades. Although, range indices are better predictors of lexical sophistication, these findings may also challenge the notion that negative correlations are associated with the selection of more advanced vocabulary that tend to be less dispersed and occur in fewer contexts. Correspondingly, the results suggest that teachers' judgments may have encompassed other writing assessment elements and criteria (e.g., task communicative achievement, organization, coherence and language) and not exclusively advanced lexical selection (e.g., the texts in Study 2 included vocabulary selection that is sufficient to respond the writing question and task). Furthermore, the analysis of the scripts using TAALES that comprises indices that are aligned to the COCA corpus might have also influenced the results. During Study 2, for example, the TAALES software analyzed high-frequency and less-frequent words that were present in the scripts by comparison with the COCA corpus. The software might have delivered results deemed as high-frequency scores and not stemming from sophisticated academic vocabulary lists. Certainly, this is important to consider, whether the lexical units selected by L2 learners in the scripts were mainly high-frequency academic words and phrases, or that teacher's judgments were based on different criteria rather than vocabulary selection exclusively, and whether teachers' high-grades encompassed a combination of both.

REVISTA VÍNCULOS UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

N-gram Word frequency and range. Only one correlation was found at the bigram association of strength, that is to say, the degree to which words are threaded and linked to one another. One measurement of that degree is the bigram Delta P (DP). In the study, the bigram DP explained only 1.88% of the variance. In addition, higher human judgments correlated with all bigram proportion indices lists indicating that less frequent bigrams in the scripts co-occurred with three bigram lists (10k to 30k) and explained a 5.28% of the variance, the bigrams in the texts co-occurred mostly with the 40k to 100k lists and explained 14.61% of the variance. This indicates that the correlation between higher grades and scripts occurred mainly at the bigram frequency proportion indices. Finally, at the trigram index group, only one correlation was found between high grades and the 40k trigram proportion list and explained only 1.04% of variance.

General Discussion

Predicting grades using automatic analyzers

The general purpose of this study was to explore the relationships between students' choices of vocabulary in formative written assignments, and the grades given by their teachers in three Ecuadorean universities. In the analysis described above, the term 'associated' in Spearman's correlations requires an emphasis on analysis of the results. Association should be interpreted as a correlation of the ranks (e.g., grades and vocabulary levels). Thus, the non-linear correlation for Study 1, for example, demonstrated that higher values found at five CEFR vocabulary levels were associated with higher grades. Furthermore, the results show that there is a higher correlation of word types at the B1 and B2 levels. In fact, these last two correlations may suggest that higher grades, previously considered in each script, validate the Text Inspector and Lexis EVP analyses. Regarding the vocabulary validations at each CEFR level, previous and similar studies (Laufer, 1994; Francois, Volodina, Pilan, & Tack, 2016; Alp, Kerge, & Pajupuu, 2013; Lenko- Szymanska, 2015) suggest that external tools, such as Text Inspector, are of value in the quest for answers to questions such as 'How many words per level should learners know'? or 'Which words at which level'? In an attempt to answer those questions, Amkham (2016) analyzed L2 writings using Text Inspector and Lexis EVP to verify the level of less frequent or 'rare' words. Amkham found that students tended to use rare words in their scripts after being taught about the word lists at each CEFR level. Similar studies show that the relationship between higher grades and the inclusion of more 'advanced' words in scripts may lead to greater lexical sophistication and writing proficiency (Laufer, 1998; Stæhr, 2008; Kyle & Crossley, 2016; Schmitt, 2005). Finally, the issue of assigning levels to individual words might be both the weakest and the strongest aspect of the Text Inspector Lexis EVP online analyzer; this tool should complement the writing grading process rather than be considered the last word in the analysis of vocabulary of L2 written texts.

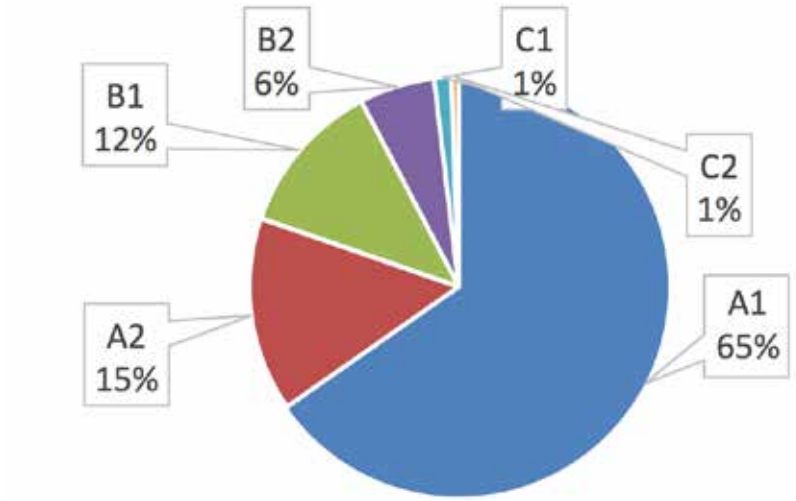
On the other hand, the construct of lexical sophistication departs from the assumption that the more difficult or advanced a word or phrase is, the less frequently it is used by a language learner. Van Gijsel, Speelman, and Geeraerts (2005) contend that measuring lexical sophistication involves a proportion of vocabulary items from 'a number of frequency bands, which are based on a (typically external) frequency list' (p. 2). The selection of TAALES (Kyle & Crossley, 2015; Allen, Crossley, & McNamara, 2015)

focuses on the ability to measure indices that can explain human judgments of lexical proficiency. In Study2, the results indices measures obtained using TAALES demonstrate that after conducting a set of Spearman's correlations between the grades and the fifty-five indices, there is a relationship with twenty-two indices of lexical sophistication (Table 6). More specifically, these results might also demonstrate that human grades in an L2 script corpus correlate with almost all indices of word frequency and range occurrences, with one bigram association of strength, all bigram proportion frequencies and only one trigram proportion frequency. Most importantly, the bigram frequency and proportion (10-100k) correlations corroborate the idea that scripts graded as having higher lexical proficiency contained and used proportionally more bigrams from the reference COCA (Kyle & Crossley, 2016, Crossley et al., 2012).

Particular aspects of lexical analyzers

Writing quality and higher grades reflect the ability of the learner not only to include less common lexical items, but also to engage with other attributes, such as the knowledge and awareness of the type of words to use in writing a composition. Grabe and Kaplan (1996) refer to these as linguistic knowledge and writing process strategies. The proficient writer, for example, is aware of the need to combine high-frequency vocabulary from A1 and A2 levels and low-frequency lexical items (e.g., B1-C2) to enhance the quality of the script. This may be associated with the interpretation of percentages in the types of words included in the original corpus (see Graph 4 below). The Text Inspector and the Lexis EVP function results superficially imply that the scripts were 65% A1, 15% A2, 12% B1, 6% B2 and 1% for C1 and C2. The correlational analysis that incorporated the raw indices and the higher grades verifies the teachers' judgments at the CEFR B1 and B2 levels.

Graph 4
EVP type of words



On the other hand, the key advantage of the TAALES lexical analyzer is its ability to simultaneously process many texts and provide information on hundreds of indices of lexical proficiency. The software, which is free to download, similarly provides an option for individual coverage information. An initial challenge for inexperienced users might involve questions such as: What to do with too many raw data scores? What do the indices actually measure? What do the indices reveal? The TAA-

REVISTA VÍNCULOS UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

LES literature seems to be influenced by the developers of the software (i.e., focused on the tool's benefits), however, there is no mention of how TAALES processes the 'flawed' vocabulary mentioned earlier in regard to Text Inspector. The tool does not include a set of indices to measure textual inconsistencies, implying that all data is measurable. Or perhaps it does reveal errors indirectly through the absence of correlational measures. The assumption might be either that the texts do not have enough academic n-grams to be supported by the COCA lists or that trigram frequencies are more common in spoken texts and less common in written contexts. In addition, in Study 2, all correlations are weak, assuming that a variety of other factors affect the overall quality of scripts. Undoubtedly, these correlational results imply causation to the extent that alternative elucidations for the correlations are challenged.

Furthermore, text-length is a factor that needs careful thought in automatic text analyses because text-length is strongly correlated with essay quality (Kyle & Crossley 2015). The original corpus comprised scripts with short texts as required at the B1 level writing production. The scripts were of several types and varied in length. Apparently, Text Inspector is better at analyzing shorter texts than TAALES, which requires longer texts for best results. Yet, the analysis of texts also depends on the goal and, in the current study, the combination of both analyzers met expectations.

CONCLUSION

The present work found a correlation between grades and scripts. In Study 1 grades were associated with script quality at the B1 and B2 CEFR levels. Similarly, in Study 2, the TAALES results broadly correlated with twenty-two indices of lexical sophistication and bigram frequency proportions were found to be as the most representative of all correlations.

Not only researchers, curriculum planners, material writers, and test developers, but teachers and students can gain access to automatic lexical analyzers. Language teachers might complement their student's assessment of writing tasks to verify CEFR vocabulary levels or engage in classroom research projects that enable students to improve their lexical skills. Teachers and students can benefit from handy vocabulary tools such as EVP and Text Inspector (Amkham, 2016).

In effect, researchers and teachers can inform institutions about new and better practices to assess L2 practices through the combination of human judgment and state-of- the-art lexical analyzers. Even though Text Inspector and TAALES are not flawless tools they can both be important in the teaching, learning, and assessment of English as a foreign language.

More advanced lexical analysis might not directly translate to the classroom, rather it might assist teachers in better understanding the intricacies of vocabulary production in writing tasks. An example of this is the fifty-five old and new indices in TAALES that allow the teacher to evaluate not only word frequency and range, but also to verify new complementary indices, such as n-gram frequency, range, association of strength, and proportional frequency lists.

Finally, TAALES relies on a number of studies that measured lexical sophistication (Kyle & Crossley, 2015, 2016). The validity of Text Inspector can be traced to studies that include specific functions of Text Inspector, for example, McCarthy and Jarvis (2010) assessed convergent, divergent, internal, and incremental validities. The results are highly technical; however, they add the element of validity necessary for research. Lastly, although both tools include validity, more research is needed to investigate various types of texts in various contexts using automatic lexical analyzers.

REFERENCES

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. Bulletin of the Psychonomic Society, 15, 147–149.

Allen, L. K., Crossley, S. A., & McNamara, D. S. (2015). Predicting misalignment between teachers' and students' essay scores using natural language processing tools. In International Conference on Artificial Intelligence in Education (pp. 529-532). Springer International Publishing.

Alp, P., Kerge, K., & Pajupuu, H. (2013). Measuring lexical proficiency in L2 creative writing. In J. Colpaert, M. Simons, A. Aerts, M. Oberhofer (Eds.), Language Testing in Europe: Time for a New Framework? (pp. 274–286). Antwerpen: Linguapolis Universiteit Antwerpen

Amkham, C. (2016). Introducing the English Vocabulary Profile (EVP) to students: An attempt at enriching students' written language. Retrieved from http://languagerese arch.cambridge.org/images/pdf/201516_Amkham_CUP_TRP_final_report.pdf

Balota, D, Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition of single-syllable words. Journal of Experimental Psychology: General, 133, 283-316.

Bax, S. (2015, January 1). Text Inspector [computer software]. Available from http://www.textinspector.com

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

British Council (2015). English in Ecuador: An examination of policy, perceptions and influencing factors. Retrieved from: https://ei.britishcouncil.org/sites/default/ files/latin-america-research/English%20in%20Ecuador.pdf

Capel, A. (2010). A1-B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. English Profile Journal 1(1): 1–11. DOI: 10.1017/S20415 36210000 048

Capel, A. (2011). The English Vocabulary Profile. Available from www.englishprofile.org

Capel, A. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. English Profile Journal 3(1): 1–14. DOI: 10.1017/S2041536212000013

Carter, R., McCarthy, M., Mark, G., & O'Keeffe, A. (2011). English grammar today: An A-Z of spoken and written grammar. Cambridge: Cambridge University Press).

Chen, Y., & Baker, P. (2014). Investigating Criterial Discourse Features across Second Language Development: Lexical Bundles in Rated Learner Essays, CEFR B1, B2 and C1. Applied Linguistics. doi:10.1093/applin/amu065

REVISTA VÍNCULOS
UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

Cohen, L., & Manion, L. (1985). Research methods in education. Croom Helm.

Council of Europe (2011). Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In McCarthy, P. M. &

Youngblood G. M., (Eds.). Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference. (pp. 214-219) Menlo Park, CA: The AAAI Press.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band- based indices of word frequency: Implications for active vocabulary research and pedagogical applications. System., 41, 965–981

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. International Journal of Continuing Engineering Education and Life-Long Learning, 21 (2/3), 170-191.

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. Language Learning & Technology, 17, 171–192.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). Predicting lexical proficiency in language learners using computational indices. Language Testing, 28, 561–580. doi:10.1177/026553221037803

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). What is lexical proficiency? Some answers from computational models of speech data. TESOL Quarterly, 45, 182–193. doi:10.5054/tq.2010.244019

Davies, Mark. (2008-). The Corpus of Contemporary American English: 425 million words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Doane, D. P., & Seward, L.E. (2011). Measuring Skewness. Journal of Statistics Education, 19(2), 1-18.

Dörnyei, Z. (2007). Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies. Oxford: Oxford University Press.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? International Review of Applied Linguistics in Language Teaching, 47, 157–177.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from Chronolex. Frontiers in Psychology, 2,1–10.

Francois, T., Volodina, E., Pilan, I., & Tack A. (2016). SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. Proceedings of LREC, Slovenia.

Grabe, W. & Kaplan, R.B. (1996). Theory and practice of writing. Harlow: Longman.

Granger, S. (2008). 'Learner corpora in foreign language education' in N. Van Deusen- Scholl and N. H. Hornberger. (eds): Encyclopedia of Language and Education Second and Foreign Language Education. Vol. 4. Springer.

Gries, S. T. (2008). "Dispersions and adjusted frequencies in corpora". International Journal of Corpus Linguistics, 13 (4), 403–437.

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next ... International Journal of Corpus Linguistics, 18(1), 137–166

Gries, S. T., & Ellis, N. C. (2015). Statistical Measures for Usage-Based Linguistics. Language Learning, 65(S1), 228-255. doi:10.1111/lang.12119

Harrison, J., & Barker, F. (2015). English Profile in practice. Cambridge, United Kingdom.: Cambridge University Press.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. English for Specific Purposes. 27 (1): 4-21.i

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. TESOL Quarterly, 49(4) 757–786. doi: 10.1002/ tesq. 194

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. Journal of Second Language Writing, 34(4), 12-24.

Kuperman, V., Stadthagen-Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. Behavior Research Methods, 44, 978–990. doi:10.3758/s13428-012-0210-4

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? RELC Journal, 25(2), 21–33. doi:10.1177/003368829402500202

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or di□erent? Applied Linguistics 19(2): 255–271. DOI: 10.1093/applin/19.2.255

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. Applied Linguistics, 16, 307–322. doi:10.1093/applin/16.3. 307

Lenko-Szymanska, A. (2015). The English Vocabulary Profile as a benchmark for assigning levels to learner corpus data. Studies in Corpus Linguistics Learner Corpora in Language Testing and Assessment, 115-140. doi:10.1075/scl.70.05len

McCarthy, P., & Jarvis S. (2010). MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods. 42(2):381–392.

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

REVISTA VÍNCULOS
UNIVERSIDAD DE LAS FUERZAS ARMADAS ESPE
PRINT: ISSN 2477-8877
ONLINE: ISSN 2631-2751

Meara, P. (1996b). The Vocabulary Knowledge Framework. retrieved from www.lognostics. co.uk/vlibrary/index.htm

Millar, E. (2016). In Search of a Common Core of Key Vocabulary among EFL Coursebooks for 4th Year Secondary Education in Cantabria Digital Lexical Notebooks for Secondary Education (Master's dissertation, Universidad de Cantabria, Santander, Spain). Retrieved from https://repositorio.unican.es/xmlui/bitstream/handle/10902/8896/MillerElaine.pdf?sequence=1&isAllowed=y

Nation, P. (1990). Teaching and learning vocabulary. Boston, MA: Heinle and Heinle. doi:10.1016/0346-251X(94)90065-5

Nation, P. (2001). Learning vocabulary in another language. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139524759

Negishi, M., Tono, Y., & Fujita, Y. (2012). A Validation Study of the CEFR Levels of Phrasal Verbs in the English Vocabulary Profile. English Profile Journal, 3. doi:10.1017/s2041536212000037

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In Proceedings of the Corpus Linguistics 2003 Conference, D. Archer, P. Rayson, A. Wilson & T. McEnery (eds), 572–581. Lancaster: University of Lancaster.

Olinghouse, N. G., & Wilson, J. (2013). The Relationship between Vocabulary and Writing Quality in Three Genres. Reading and Writing: An Interdisciplinary Journal, 26, 45-65.

Richards, J. (1976). The Role of Vocabulary Teaching. TESOL Quarterly, 10(1), 77. doi:10.2307/3585941

Saville, N. & Hawkey, R. (2010). The English Profile Programme - the first three years, English Profile Journal 1 (1).

Saville, N. (2012). The English Profile: Using Learner Data to Develop the CEFR for English. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), Developmental and Crosslinguistic Perspectives in Learner Corpus Research (pp. 17–26). Amsterdam & Philadelphia: John Benjamins.

Schmitt, N. (1995). A Fresh Approach to Vocabulary: Using a Word Knowledge Framework. RELC Journal, 26(1), 86-94. doi:10.1177/003368829502600105

Schmitt, N. (1998). Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. Language Learning, 48(2), 281-317. doi:10.1111/1467-9922.00042

Schmitt, N. (2005). Lexical resources in Main Suite writing examinations, in Lim, G. & Galaczi, D. (2010). Lexis in the assessment of speaking and writing: An illustration from Cambridge ESOL's General English tests. Research Notes 41, 14-19, Cambridge: Cambridge ESOL. Retrieved from www.CambridgeESOL.org/ rs_notes/offprints/pdfs/RN41p2-7.pdf

Schmitt, N. (2012). Vocabulary in language teaching. New York: Cambridge University Press.

Schmitt, N. & Meara, P. (1997). 'Researching vocabulary through

a word knowledge framework. Word associations and verbal suffixes.' Studies in Second Language Acquisition 20: 17-36.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. Applied Linguistics, 31, 487–512. doi:10.1093/applin/ amp058

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. Language Learning Journal 36: 139–152. DOI: 10.1080/09571730802389975 Stubbs, M. (2007). Quantitative data on multi-word sequences in English: The case of word 'world'. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), Text, Discourse and Corpora: Theory and Analysis (pp. 163–189). London: Continuum.

van Ek, J. (1980). Threshold Level English. Oxford: Pergamon Press.

van Gijsel, Speelman, D. & Geeraerts, D. (2005) A variationist, corpus linguistic analysis of lexical richness. Proceedings from the Corpus Linguistics Conference Series 1 (1), p. 1-16.

BIOGRAFÍA DE LOS AUTORES

José Lema Alarcón

Formación Académica: estudió en Inglaterra, en donde obtuvo su título de Ph.D., y maestría en Investigación Aplicada en la Universidad de Exeter, así como también una maestría en la Enseñanza del Idioma Inglés TESOL por parte de la Universidad de Canterbury Christ Church. Su trabajo se enfoca en la investigación de corpus lingüístico que permita el análisis de textos a gran escala. Para lo cual el Dr. Lema utiliza herramientas de minería de datos o Data Mining, Big Data y modelos estadísticos de predicción. También, el Dr. Lema realiza investigación relacionada con el aprendizaje de una segunda lengua a través del uso de la tecnología en la educación (ej., blended learning, plataformas académicas, inteligencia artificial).