Aspects of EFL University Learners' Lexical and Phraseological Proficiency as Predictors of Writing Quality

Brahim Ait Hammou
Ministry of National Education

Mohammed Larouz
Moulay Ismail University

Mustapha Fagroud
Moulay Ismail University

Fouad Akki
Moulay Ismail University

Abstract

This study aims to examine the relationship between the productive knowledge of some lexical and phraseological indices and the quality of English as a Foreign Language (EFL) learners' writing. A sample of 120 expository essays, written by semesters 1 and 5 university students in a less proficient EFL context, are rated by human evaluators and automatically examined for the target indices. The results show that, unlike the index of lexical diversity, both indices of content word frequency and range could significantly discriminate between different proficiency levels. For the phraseological indices, both the proportions of rare and frequent bigrams yielded between-group differences, with higher proficiency students performing significantly better in both categories. Using a regression analysis, the results show that the use of rare and contextually restricted content words and the production of larger proportions of rare and frequent bigrams could be considered indicators of better writing proficiency. The study suggests implications for the teaching of EFL.

Résumé

Cette étude vise à examiner la relation entre les compétences lexicales/phraséologiques et la qualité de la production écrite des étudiants de l'anglais/langue étrangère. Un échantillon de 120 essais explicatifs écrits par des étudiants des semestres 1 et 5 dans un contexte EFL moins compétent, sont notés par des évaluateurs humains et examinés pour les indices cibles en utilisant un traitement automatique. Les résultats montrent que, contrairement à l'indice de diversité lexicale, la fréquence et la distribution contextuelle de mots de contenu peuvent distinguer entre les deux différents niveaux. Pour les indices phraséologiques, les proportions de bigrammes rares et fréquents ont produit des différences entre les deux groupes : les étudiants de S5 sont les plus performants avec de meilleurs résultats dans les deux catégories. À l'aide d'une régression linéaire, les résultats montrent que l'utilisation de mots de contenu rares et contextuellement restreints et aussi une plus grande proportion de bigrammes rares et fréquents pourraient être considérés comme des indicateurs d'une meilleure compétence

en production écrite. L'étude suggère des implications pour l'enseignement et l'apprentissage de l'anglais en tant que langue étrangère.

Aspects of EFL University Learners' Lexical and Phraseological Proficiency as Predictors of Writing Quality

Introduction

The development of learners' lexical knowledge at the level of both lexical sophistication and phraseological knowledge has gained paramount importance with advances in computational linguistics. Recent research in phraseology has highlighted that "language is highly patterned" (Römer, 2009, p. 141). This principle is emphasized by Sinclair (1991, 1996), who maintains that a large amount of a speaker's knowledge is composed of ready-made linguistic patterns. Cognitive linguistic research (e.g., Ellis, 2002a, 2002b) has underlined the role of frequency in language development. Previous studies (e.g., Durrant & Schmitt, 2009; Laufer,1994; Laufer & Nation, 1995) show that learners' lexical knowledge is to a large extent influenced by its frequency. Contextual distribution, or word range, also impacts lexical production in the sense that more proficient learners produce texts with more contextually restricted lexical items (Crossley et al., 2013; Gries, 2010; Kyle & Crossley, 2016). Similarly, calculating the diversity of vocabulary in a text has also been facilitated by the availability of sophisticated computer programs. Lexical diversity (henceforth, LD) in learner corpora has been reported to be an indicator of learners' vocabulary size and also to affect the perceived quality of their writing (e.g., Crossley & McNamara, 2012; González, 2017; McNamara et al. 2010; Vidal & Jarvis, 2018; Zenker & Kyle, 2021).

The current study aims at exploring the productive use of lexical and phraseological indices and their predictive ability of the quality of EFL learners' writing in a context where no similar study has been conducted before. The current study explores the relationship between the target indices and learners' writing quality in a context where the teaching of English as a foreign language starts only late in the secondary school. In the Moroccan context, starting from their early primary school years, learners are introduced to Standard Arabic and French, neither of which is a first language (L1), while English is introduced to them later in the third year of the secondary school.

Literature Review

Lexical Sophistication

Read (2000) defines lexical sophistication as a measure of the proportion of advanced words in a text. One of the first attempts to profile learners' vocabulary based on its frequency is Laufer and Nation's (1995) Lexical Frequency Profile (LFP). This technique categorizes learners' vocabulary into frequency bands. Recent approaches to measuring lexical sophistication go beyond classifying the words into bands, and they adopt a count-based procedure which assigns every word in a text a frequency score based on its occurrence in a larger reference corpus such as COCA or the BNC.

As Schmitt (2010) maintains "the frequency in which a word occurs in language permeates all aspects of vocabulary behavior. It is arguably the single most important

characteristic of lexis that researchers must address" (p. 63). Daller et al. (2007) concluded that measures of lexical sophistication correlated significantly with teachers' ratings of students' essays. In their study of the measures which affect the quality of students' writing as judged by independent raters, McNamara et al. (2010) noted that word frequency showed the largest difference between high and low-proficiency essays. Comparing L1 and second language (L2) learners, the authors reported that L2 learners produced texts which were characterized by the use of more frequent words compared to their L1 peers. The study concluded that "high-proficiency writers use words that occur less frequently in language" (p. 70). Jung et al. (2019) also noted that the frequency of content words is a strong predictor of writing quality. These results showed that more proficient second language learners produce texts with more sophisticated words. This finding is also documented in Crossley, et al. (2010) who showed that word frequency correlated significantly with the writing score (r = .61).

However, the role of frequency as an indicator of better language proficiency is not consistent across studies. Guo et al. (2013) explored the ability of lexical sophistication, cohesion and syntactic indices to predict second language writing proficiency in TOEFL iBT integrated and independent writing tasks. The study reported that content word frequency correlated with the students' scores in integrated essays (r = -.436, p < .001). Similarly, content word frequency correlated with the scores in the independent essays (r = -.295, p < .001). However, content word frequency alone added only 1% to the variance in integrated essays and this was after entering six other (syntactic, lexical and cohesion) variables into the regression analysis. Similarly, González (2017) reported that monolingual advanced English writers produced significantly lower frequency words compared to multilingual students enrolled in a university program. However, when intergroup differences were examined, the index of word frequency used in the study could significantly show differences between different multilingual student groups.

Word Range

Relying on the frequency of lexical items alone as an indicator of language proficiency has been an established practice. Gries (2008) maintains that "the most frequently used statistic in corpus linguistics is the frequency of occurrence of some linguistic variable. […] However, even though this is apparently not recognized much in the field, frequencies of (co-)occurrence may sometimes be incredibly misleading" (p. 403-404). To address this issue, researchers have suggested using range as a measure which is likely to balance the effects of word frequency.

Monteiro et al. (2018) maintain that range is "a measure that indicates the number of unique contexts in which linguistic items appear"(p. 4). Hence, range is a count of the different contexts rather than the different times in which a word appears in the corpus. Kyle and Crossley (2015, 2016) maintain that words with a limited range occur in a limited number of contexts while words which are contextually dispersed are associated with larger range values, indicating that they are less sophisticated.

Examining lexical sophistication in argumentative writing, Kyle and Crossley (2016) reported a significant relationship between word range and essay quality. In their study, word range explained 17% of the variance in essay scores. Also, Kyle and Crossley (2015) noted that the index of range alone explained 26% of the variance in ESL students' writing scores. In a subsequent study, the researchers reported that range could explain 17%

of variance in the writing score (Kyle & Crossley, 2016). Monteiro et al. (2018) investigated the variables which contribute to variance in TOEFL scores among L2 learners. They reported that the range index explained 17% of this variance. Taken together, these studies highlight the importance of range as a lexical indicator of language proficiency.

Lexical Diversity

LD has gained a lot of attention in studies of lexical sophistication. LD highlights the ratio of new words, that is, *types*, compared to word repetitions or *tokens* in a text. Jarvis (2013) maintains that LD is the opposite of repetition. In a lexically diverse text, there is less repetition and there are more new word types instead. McCarthy and Jarvis (2010) refer to this aspect as "the range of different words used in a text, with a greater range indicating a higher diversity" (p. 381). Jarvis and Hashimoto (2021) maintain that diversity is the inverse of repetition.

Various measures have been introduced to compute LD in a text. Early measures include the type-token ratio (Johnson, 1944), D and HD-D (McCarthy & Jarvis, 2007, 2010). However, various studies reported that these measures are affected by text length (Malvern et al., 2004; McCarthy & Jarvis, 2010; Treffers-Daller, 2013). In a study comparing a variety of indices, including D, MTLD (the Measure of Textual Lexical Diversity; McCarthy, 2005) and MATTR (the Moving Average Type-Token Ratio; Covington & McFall, 2010), Fergadiotis et al. (2015) concluded that MTLD and MATTR are the strongest measures of LD. A more recent measure of LD is MTLD-W (MTLD Wrap Around; Vidal & Jarvis, 2018). Vidal and Jarvis (2018) examined the writings of Spanish university students of English. Their results showed that MTLD and MTLD-W were affected by differences in text-length while MATTR showed consistent results. The researchers concluded that after three years of studying English at university, learners did not improve their vocabulary in a way which would result in lexically diverse texts. More satisfying results about the stability of MATTR as a measure of lexical diversity across different text lengths were reported by Zenker and Kyle (2021). The researchers examined the minimum text length needed to produce a stable LD measure, and they concluded that MATTR showed the lowest correlations with text length. The results of the study "indicate that the most stable index of lexical diversity is MATTR" (Zenker & Kyle, 2021, p.12).

Various studies examined the link between LD and writing proficiency. Crossley, Kyle et al. (2014) examined the relationship between a variety of linguistic micro-features and the quality of students' writings using a corpus of independent TOEFL essays written by Hong Kong high school students. The results showed that essays scored better by human raters contained more word types, suggesting that LD is a feature of better writing. The study reported that the index of the number of types explained 43% of the variance in essay scores. In a similar study, McNamara et al. (2010) concluded that MTLD showed the largest effect size between high and low proficiency essays. This index correlated with essay scores (r = -.35, p < .001) and it contributed to predicting variance in students' scores. Using a corpus of essays written by graduating Korean high school students, Crossley and McNamara (2012) reported that the D index of LD accounted for 18% of the variance in the entire essay corpus scores alone, and it showed a correlation of r = .427 with the essay scores. Similarly, González (2017) reported that MTLD showed the greatest difference between the essays written by monolingual and those written by multilingual

English writers. Using essays written by ESL students, Jansen et al. (2021) showed that raters judged texts with greater diversity more positively both holistically and analytically. Using different measures of LD and also different operationalizations of word types, Jarvis and Hashimoto (2021) examined LD in essays written by native and non-native English speakers. The study reported significant correlations between human ratings of LD and all automated measures used, although MTLD, followed by MATTR, showed significant correlations with most measures of LD. For MATTR, which is adopted in the current study, Jarvis and Hashimoto (2021) reported that operationalizing types as automatically-generated lemmas yielded better correlations with human scores of LD (r = .501) compared to orthographic forms (r = .499), flemmas, and word families. Other studies (Crossley, Kyle, et al. (2014); Crossley, Salsbury, & McNamara, 2014; Jarvis, 2017; Treffers-Daller, 2013; Vidal & Jarvis, 2018) reported similar conclusions about the importance of LD in judging writing quality, although they relied on different measures.

Phraseological Knowledge

Phraseology has gained a prominent position in studies of language learning since the early works of Firth (1957) and then of his followers such as Sinclair (1991, 1996) and Halliday (1966) and also other researchers who followed the Russian school in phraseology such as Cowie (e.g., 1994, 1998), Howarth (e.g., 1996, 1998) and Mel'cuk (1995, 1998). With the sophistication of computer tools, it has become possible to examine a variety of phraseological phenomenon in huge corpora. Phraseology is concerned with the analysis of formulaic language, which Wray (2000) defines as "a sequence, continuous or discontinuous, of words or other meaning elements, which appears to be prefabricated: that is stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (p. 465). Similarly, Gries (2008) defines phraseologisms "as the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance" (p. 3). Within this frequency-based perspective, Sinclair (1987) maintains that a collocation is a significant word co-occurrence which occurs with "a greater frequency than expected by chance" (p. 70). Halliday and Hassan (1976, as cited in Hoey, 2005) define a collocation as a "a cover term for the kind of cohesion that results from the co-occurrence of lexical items that are in some way or other typically associated with one another, because they tend to occur in similar environments" (p. 287). The present study is concerned only with contiguous two-word sequences, labelled bigrams. Usually, researchers consider a bigram, or an n-gram of any length, any contiguous words whose co-occurrence is beyond the effect of chance (Paquot & Granger, 2012).

Studies which examine word combinations mainly use two strength of association measures: The MI (the Pointwise Mutual Information) and the *t*-score statistics. Strength of association measures are used to test the existence of an association between the two words (the *t*-score) and the strength of this association (MI score) (Durrant & Schmitt, 2009; Evert, 2005, 2009). Durrant (2014) maintains that "MI is a measure of the extent to which the probability of meeting one word increases once we encounter the other" (p. 456). Qian (2019) believes that "the MI-value measures how strongly words are attracted to each other. A high MI is indicative of a collocation that is idiomatic or of high quality" (p. 3). Durrant and Schmitt (2009) maintain that the MI score highlights collocations made up of words

that are rarely found independently of each other. The authors maintain that "MI tends to give prominence to word pairs which may be less common, but whose component words are not often found apart" (p. 167). Higher MI scores indicate that there is stronger attraction between the words of the collocation. Gablasova et al. (2017) describe MI statistic as highlighting rare exclusivity. That is, MI highlights collocations composed of words which are strongly associated but rare in the corpus, such as *exultant triumph* (Evert, 2009; Gries & Ellis, 2015; Gries & Stefanowitsch, 2003; Kim et al., 2018). MI statistic is calculated based on the expected (E) and observed (O) frequencies of the co-occurring words in a reference corpus. The formula is: $MI = \log 2 \frac{O}{E}$ (Evert, 2009).

The second measure of strength of association is the *t*-score. Durrant (2014) says that "*T*-score […] is a hypothesis testing technique, which evaluates how much evidence there is that a particular combination occurs more frequently than we would expect by chance alone" (p. 456). The top scoring bigrams identified by the *t*-score are composed mostly of grammatical words and high frequency lexical verbs such as *think*, *want*, *get*, *say* (e.g., Bestgen, 2016a, 2016b, 2019; Bestgen & Granger, 2014; Granger & Bestgen, 2014, 2017). This measure is calculated as $t\text{-}score = \frac{O-E}{\sqrt{O}}$, with O being the observed and E the expected frequencies of the co-occurring words in the reference corpus. The *t*-score measures the degree of confidence we can have about the existence of association between two words. Compared to MI, the *t*-score is a measure of frequency highlighting the number of times a collocation is observed in the corpus rather than exclusivity. Granger and Bestgen (2017) maintain that "the advantage of using two association scores is that they bring out word combinations of a different nature: MI tends to highlight word sequences made up of low-frequency words […], while the *t*-score singles out those composed of high-frequency words"(p. 388). An MI score which equals or is greater than 3 (MI ≥ 3) and a *t*-score which equals or is greater than 2 (*t*-score ≥ 2) are thresholds for a word combination to be strong (e.g., Bestgen, 2016a; Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Evert, 2005, 2009).

Empirically, Durrant and Schmitt (2009) examined the use of collocations in essays written by post-graduate students on pre-sessional EAP courses at a British university and first year undergraduate students on in-sessional EAP courses at an English-medium university in Turkey. The study concluded that native writers use larger proportions of lower-frequency, strongly associated collocations (highlighted by MI) than non-native writers who rely more on high-frequency collocations, highlighted by the *t*-score. Similarly, Durrant (2014) concluded that native writers produced more infrequent strongly associated collocations compared to non-native ones. Granger and Bestgen (2014) concluded that advanced non-native learners overuse high-frequency collocations and underuse strongly-associated lower-frequency collocations. The authors concluded that "L2 learners' failure to use native-like formulaic sequence is one factor in making their writing feel non-native" (p. 86). Wolk et al. (2017) examined stories written by L2 writers, and they concluded that intermediate texts are characterized by using a lower proportion of less frequent but strongly associated sequences and a higher proportion of high-frequency sequences.

Assessing writing quality, Bestgen (2016a) showed that formulaic measures are the best predictors of writing quality, compared to single-word lexical measures. The researcher noted that the mean MI score is the most effective bigram measure which has a strong relationship with the raters' judgements of text quality. The study highlights that the top-scoring bigrams which are identified by MI are composed of rare words. This suggests

that rare bigrams are taken as a better indicator of writing quality. A similar conclusion is outlined in Bestgen and Granger (2014), Granger and Bestgen (2017), and also Wolk et al. (2017). These studies concluded that while the mean *t*-score of the bigrams produced by learners never showed a statistically significant relationship with the writing score, the mean MI score was significantly linked to the quality of the text as judged by independent raters. Crossley et al. (2012) also observed that essays which are scored better by expert raters contain fewer frequent n-grams while they contain a larger proportion of rare n-grams. The authors noted that knowledge of n-grams, compared to other lexical indices, is a significant predictor of essay quality. A similar finding was highlighted by Kyle and Crossley (2016) who concluded that, compared to word frequency, bigram frequency is an important indicator of writing quality.

Although most studies seem to indicate that second language learners learn and produce highly frequent combinations earlier while rare and strongly associated ones "take longer to acquire" (Durrant & Schmitt, 2009, p. 175), not all studies adhere to this conclusion. For instance, Yoon (2018) examined the development of ESL students' writing proficiency in the argumentative and narrative writings using a variety of measures, including phraseological ones. The study highlighted that for both types of writing, the use of frequent n-grams contributed significantly to the quality of students' writing. A similar conclusion was reported earlier by Kyle and Crossley (2015) who showed that the use of frequent trigrams correlated significantly with the quality of students' scores in writing.

Methodology

Rationale

This study has two main objectives: First, it aims to examine cross-sectional lexical and phraseological differences between two groups of EFL university students: Semester 1 (S1) students, who are university freshmen, and Semester (S5) students. Second, the study has the objective of identifying which among the examined indices contribute(s) to EFL students' writing quality as perceived by human judges. Although many studies with similar objectives have been published in other contexts using quite advanced learner corpora, no similar study has been carried out in the Moroccan context where the teaching of English starts only late in the secondary school. Therefore, examining the cross-sectional changes of aspects of lexical and phraseological indices and their relationship with the writing quality using less advanced learner data might inform language learning research by providing evidence from an unexplored context. In the current study, proficiency is taken as equivalent to grade level. This study is guided by the following research questions:

1. Does the amount of time spent studying English yield significant differences between university S1 and S5 undergraduates at the level of lexical sophistication, operationalized as average word frequency and range in a reference corpus, and lexical diversity as reflected through their writing?
2. Does the amount of time spent studying English yield significant differences between S1 and S5 at the level of their use of bigrams in their writing?
3. Which among these target *lexical* and *phraseological* indices correlate with and could predict the quality of students' essays as reflected from human judgement?

Data Collection and Sampling

A sample of 120 participants, majoring in English (EFL), was randomly selected from an initially available data set of essays written by a total number of 167 S1 and 174 S5 students respectively. Due to constraints related to finding raters who could score the whole data set, we limited this study only to 60 participants from each group. This random sampling was done using SPSS random set selection functionality.

With regards to data collection, a writing test was administered at the beginning of the academic year (October). The participants from the target groups were asked to write a timed and invigilated essay of approximately 300 words during their officially-scheduled classes. The informants were required to write their essays in response to a unified expository prompt, and they were given a maximum of one hour to complete the essay. The writing prompt was as follows:

> *Young people study because they want to have a job when they grow up. What are the other reasons for which young people go to school?*

Table 1

The Number of Word Types and Tokens in the Participants' Essays

Proficiency Level		N	Min.	Max.	Mean	Std. Dev.
S1	tokens	60	81	265	146.37	42.25
	types		36	127	73.36	17.98
S5	tokens		98	309	172.62	45.93
	types		58	152	88.01	19.18

Data Processing and Analysis

Before processing the data, the learner corpus was edited in the following ways. For the texts used in lexical analysis, capitalization, spelling and minor grammar mistakes were corrected. We also changed the abbreviations and the contractions to their complete forms. In the texts which were used for bigram analysis, errors made in word combinations were left as they were. Finally, since this study used COCA as its reference corpus, we normalized the texts to the American English spelling.

For the analysis of word frequency and range, the corpus was processed using the open-source computer tool TAALeS (version 2.0) (Crossley & Kyle, 2018; Kyle & Crossley, 2015; Kyle et al., 2018). For this study COCA academic word frequency and range indices were examined. We opted for using only the academic sub-section of the COCA corpus as our reference corpus because we think that our data would be widely influenced by aspects of academic language since our informants (S1and S5 students) study writing in formal classes which heavily focus on the academic aspects and conventions of writing.

TAALeS generates frequency scores for each word token. It calculates an average frequency mean for each text by dividing the sum of the frequency values by the number of the tokens that received a score in the text. TAALeS was also used to compute the range index. Each word token in a text is given a range value based on the number of texts in which it appears in the academic section of COCA. In this study, word frequency and range

were computed both for content and function words using the automatically generated log-transformed values.

For the analysis of LD, the computer program TAALED (version 1.4.1) (Kyle et al., 2020; Zenker & Kyle, 2021) was used. To compute LD, the present study adopted MATTR, a 50-word window measure (Covington & McFall, 2010), which has been reported to be more reliable in measuring LD (Fergadiotis et al., 2015; Fergadiotis et al., 2013; Jarvis & Hashimoto, 2021; Kyle & Crossley, 2015; Kyle et al., 2020; Vidal & Jarvis, 2018; Zenker & Kyle, 2021). Jarvis (personal communication, July 4th, 2019) also recommended using MATTR stating that, compared to other measures, "MATTR might be slightly an even better measure of lexical diversity than MTLD and MTLD-W." In the present study, LD is measured in learner texts based on uncorrected, automatically-generated word lemmas. A lemma includes a headword and its most immediate inflections within the same part of speech category. TAALED considers word types as instances of the same lemma (Kyle et al., 2020). For example, the words *speak*, *speaks*, *spoke*, *spoken* are counted as one *type* or one lemma *speak*. Jarvis and Hashimoto (2021) reported that using automatically-generated lemmas in the analysis of LD produced the highest correlation between LD scores calculated by MATTR and human ratings of LD (r = .501).

For bigram analysis, the present study adopted two widely used strength of association measures to determine the collocational power of a bigram, namely the MI and *t*-score statistics. The MI and the *t*-score statistics are computed for bigram *types* based on lemmatized word forms. Following previous research on collocations (e.g., Granger & Bestgen, 2014; Bestgen, 2016a, 2016b; Durrant & Schmitt, 2009), we adopted an MI score of 3 or above and a *t*-score of 2 or above for a collocation to be associated beyond chance. Bigram analysis was carried out using the open-source COCA Parser tool (Wolk et al., 2017). COCA Parser is a web tool which takes in input texts from the user, uses CLAWS to part-of-speech tag them, and then analyze a variety of n-grams using (the lemmatized) COCA as its reference corpus. The tool calculates the MI and *t*-score for each n-gram *type* based on their appearance in the reference corpus. MI highlights n-grams composed of words of lower frequency, but which are strongly attached to each other, while the *t*-score highlights n-grams composed of high-frequency combinations. Together the measures capture how word combinations appear in a text based on the frequency of their composing words.

To deal with the third research question, which is concerned with the ability of the examined indices to predict the quality of students' scores in writing, a Pearson-Product-Moment correlation and a multiple linear regression analysis were conducted after checking that the assumptions were met. For this purpose, we set an $r \geq .70$ for two indices to be multicollinear. The correlational analysis showed that the indices of the frequency of content words and the range of content words correlated at r = .95, $p < .001$. Therefore, we discarded the index of content word frequency from the regression analysis as it showed a comparatively lower correlation with the writing score. We also made sure that the data for the dependent variable (the holistic writing score) was normally distributed as shown both from the Shapiro-Wilk test (W = .985, df = 120, p >.05) and the residuals histograms. Also, the P-P plots showed that there were no outliers in the data. We ensured that there were no issues of multicollinearity (Table 5). The analysis showed that all the values are above .10 for tolerance and less than 10 for variance inflation values (VIF). Only the variables which

correlated significantly ($p < .05$) and meaningfully ($r > .1$) with the writing score were included in the model. Hence, the index of LD was not included in the regression model.

Results

Examining Students' Lexical Production

Table 2 presents the descriptive statistics related to the performance of the two groups in relation to single word lexical indices: word frequency, word range, and LD.

Table 2
Group Means for Word Frequency, Word Range and Lexical Diversity

Descriptive Statistics Grade Level		Word frequency (log)		Word range (log)		Lexical diversity
		content words	function words	content words	function words	all words
S1	N	60				
	Mean	2.55	3.89	-.44	-.04	.73
	Std. Deviation	.09	.13	.06	.02	.04
	Min.	2.34	3.59	-.58	-.10	.63
	Max.	2.84	4.18	-.30	-.01	.82
S5	N	60				
	Mean	2.51	3.89	-.47	-.04	.74
	Std. Deviation	.089	.09	.05	.02	.04
	Min.	2.25	3.64	-.66	-.11	.65
	Max.	2.69	4.14	-.36	-.02	.84

Table 2 indicates that there are some differences between the two grades at the level of content words both when we consider word frequency and word range. For content word frequency, the difference is statistically significant ($t = 2.55$, $df = 118$, $p < .05$, $d = .43$) with a small effect size (Cohen, 1969). For the range of content words, the difference between the two groups is also statistically significant with a medium effect size ($t = 3.23$, $df = 118$, $p < .05$, $d = .50$). Unlike content words, the frequency and range of function words showed very similar group means and no significant difference was observed between the two groups ($p > .05$).

For LD, the descriptive statistics indicate that there is no large difference between the two groups in terms of the variation of their vocabulary. The MATTR mean ratio for S1 essays is .73 which increased only by one point to .74 for S5. The independent-samples t-test showed that this small difference is not significant ($t = -1.21$, $df = 118$, $p > .05$), which indicates that after studying English for two years at university, S5 learners produced texts which are similar to those of university freshmen at the level of the diversity of their produced texts.

The Development of Learners' Production of Bigrams

This section presents results related to learners' productive use of bigrams. We tested the hypothesis that the two proficiency groups are not significantly different in their production of rare and frequent bigrams as highlighted by the MI and *t*-score statistics respectively. Table 3 provides descriptive statistics related to the proportions of MI and *t*-score bigrams. Examples of the highest scoring bigrams which are highlighted by the two statistics are provided in the Appendix.

Table 3

Descriptive statistics: Proportions of the produced rare and frequent bigrams

Proficiency Level		proportion of MI bigrams	proportion of *t*-score bigrams
S1	N	60	
	Mean	27.93	72.91
	Std. Deviation	7.61	8.01
	Min.	12.82	35.26
	Max.	69.90	90.00
S5	N	60	
	Mean	30.81	77.00
	Std. Deviation	4.82	6.79
	Min.	22.45	39.67
	Max.	45.38	91.18

Table 3 shows that there is an increase in the proportions of both categories of bigrams (those highlighted by MI and those highlighted by the *t*-score) when we compare S1 to S5 students. For the bigrams which are highlighted by MI, S1 had a mean proportion of 27.93 which increased to 30.81 for S5. For the bigrams highlighted by the *t*-score, the results indicate that there is also an increase in the proportion of this category from a mean of 72.91 for S1 students to 77.00 for S5.

Using an independent-samples *t*-test, the results show that the difference between the two groups in the category of MI bigrams is statistically significant with a small effect size ($t = -2.47$, $df = 118$, $p < .05$, $d = .45$). A similar result was found for the category of bigrams with the *t*-score ($t = -3.01$, $df = 118$, $p < .05$, $d = .55$), though the difference is moderate. By comparing the performance of S1 to S5 students, it seems that S5's productive knowledge of bigrams significantly improves over the first two years at university towards the production of a larger proportion both of rare (MI) and frequent (*t*-score) units.

The relationship between the lexical/phraseological indices and the writing score

A Pearson-product-moment correlational analysis and a multiple linear regression were conducted to examine the relationship between the target lexical-phraseological indices and students' holistic writing scores. The correlations are displayed in Table 4.

Table 4

Correlations between lexical/phraseological indices and the writing score

Pearson-Product-Moment Correlations										
		Mean writing score	Token count	frequency of content words (log)	frequency of function words (log)	Range of content words (log)	Range of function words (log)	Lexical diversity (MATTR)	Proportion of MI bigrams	Proportion of *t*-score bigrams
Mean writing score	Pearson Correlation	1								
Token Count	Pearson Correlation	.422*	1							
	Sig. (2-tailed)	.000								
Frequency content words (log)	Pearson Correlation	-.209*	-.093	1						
	Sig. (2-tailed)	.022	.312							
Frequency function words (log)	Pearson Correlation	-.177	-.086	-.032	1					
	Sig. (2-tailed)	.053	.349	.731						
Range of content words (log)	Pearson Correlation	-.290*	-.101	.950*	.002	1				
	Sig. (2-tailed)	.001	.272	.000	.981					
Range of function words (log)	Pearson Correlation	-.198*	-.169	.065	.654*	.089	1			
	Sig. (2-tailed)	.030	.064	.484	.000	.334				
Lexical diversity (MATTR)	Pearson Correlation	.095	.101	-.228*	-.014	-.253*	-.008	1		
	Sig. (2-tailed)	.302	.270	.012	.881	.005	.931			
Proportion of MI bigrams	Pearson Correlation	.267*	.070	-.188*	.075	-.202*	.083	.099	1	
	Sig. (2-tailed)	.003	.447	.040	.418	.027	.367	.281		
Proportion of *t*-score bigrams	Pearson Correlation	.345*	.108	-.045	.046	-.064	.035	-.115	.401*	1
	Sig. (2-tailed)	.000	.242	.624	.618	.486	.706	.213	.000	
	N**	120								

The significant correlations are flagged with *

N** = 120 for both groups in all the variables

Prior to presenting the data pertaining to correlations and regression, it is important to note that the difference between the two grade levels in their mean writing scores is statistically significant (S1 Mean = 11.96, *SD* = 1.35; S5 Mean = 13.93, *SD* = 1.19; $t = -8.45$, $p < .01$, $d = 1.54$) (the adopted scoring scale is from 0 to 20), which might reflect the difference in their proficiency as operationalized by grade level.

For the correlations between the target indices and the holistic writing score, it is important to note that we added to the correlations the index of the overall token count (i.e. text length) because we wanted to make sure that the writing scores are related to the target indices rather than to text length. The results of the correlations (Table 4) show that number of tokens significantly correlated with the writing score. However, none of the target lexical or phraseological indices correlated significantly with the number of tokens. This indicates that any correlations between the target indices in this study and the holistic writing score are not related to text length but to the target lexical or phraseological index itself.

The index of the proportion of frequent bigrams (highlighted by *t*-score) showed the highest correlation with the writing score ($r = .345$, $p < .001$). Similarly, the index of the proportion of rare bigrams (highlighted by MI) showed a significant, small effect size, correlation with the writing score ($r = .267$, $p < .001$). This indicates that the use of higher proportions both of frequent and rare bigrams increases the chances of obtaining better writing scores.

For the single word lexical measures, the frequency of content words showed a small negative correlation with the writing score ($r = -.209$, $p < .05$), which indicates that the production of texts with more sophisticated (i.e. less frequent) vocabulary might increase the perceived quality of an essay. The index of the frequency of function words showed a negative but non-significant correlation with the writing score ($r = -.117$, $p > .05$). Finally, the index of range for both content ($r = -.290$, $p < .01$) and function ($r = -.198$, $p < .05$) words showed a significantly negative but small effect size correlation with the writing score, suggesting that the production of texts with contextually restricted vocabulary is likely to enhance the quality of one's writing. The index of LD did not show any significant correlation with the writing score ($r = .095$, $p > .05$).

Before conducting the regression analysis, the index of the frequency of content words was removed because it showed a very high correlation with the index of the range of content words ($r = .95$, $p < .001$), which we decided to keep in the regression model because of its better correlation with the writing score.

Table 5

Collinearity Diagnostics

Collinearity Diagnostics[a]								
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Range of content words (log)	Range of function words (log)	Proportion of MI bigrams	Proportion of *t*-score bigrams
1	1	4.79	1.00	.00	.00	.01	.00	.00
	2	.15	5.55	.00	.00	.91	.02	.00
	3	.02	12.86	.03	.10	.08	.86	.01
	4	.01	18.92	.04	.70	.00	.05	.23
	5	.00	33.78	.93	.20	.01	.06	.76
a Dependent Variable: Mean writing score								

The results of the regression analysis indicate that the model is statistically significant: ($F_{(4, 115)} = 8.97$, $p < .05$). As Table 6 shows, the four variables entered into the regression model could account for 23% of the variance in the holistic writing scores ($R = .48$, $R^2 = .23$), which suggests that the four target indices could significantly be taken as predictors of writing quality.

Table 6

Summary of Multiple Linear Regression Model

Model Summary[b]							
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics		
					df1	df2	Sig. F Change
1	.48[a]	.23	.21	1.43	4	115	.000
[a] Predictors: (Constant), proportion of t-score bigrams, range of function words (log), range of content words (log), proportion of MI bigrams							
[b] Dependent Variable: Mean writing score							

The Beta and *t*-test analysis (Table 7) showed that three of the four variables in the model contributed significantly to the prediction of the writing score. The first index which showed the highest standardized Beta score is the proportion of frequent bigrams (S. Beta = .28, $t = 3.24$, $p < .05$). However, the proportion of MI bigrams didn't show any significant contribution to predicting variance in the writing scores (S. Beta = .12, $t = 1.33$, $p > .05$). The range of content words index contributed significantly to the model with a large Beta value (S. Beta = -.22, $t = -2.74$, $p < .05$) suggesting that this index is an important contributor to predicting variance in writing scores. Finally, the range of function words also showed a significant Beta value with the writing score (S. Beta = -.19, $t = -2.40$, $p < .05$) which indicates that the production of function words with a limited range might significantly contribute to the quality of one's writing, although this index didn't reflect between-group differences.

Table 7

Summary of the Coefficient Scores

Coefficients[a]											
Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	4.12	1.58		2.60	.01					
	range content words log	-5.94	2.16	-.22	-2.74	.00	-.29	-.24	-.22	.94	1.05
	range function words log	-15.06	6.26	-.19	-2.40	.01	-.19	-.21	-.19	.98	1.01
	Proportion of rare bigrams *(MI*)	.03	.02	.12	1.33	.18	.26	.12	.10	.80	1.24
	Proportion of frequent bigrams (*t*-score)	.06	.01	.28	3.24	.00	.34	.29	.26	.83	1.19

a. Dependent Variable: Average score in writing

Discussion and Conclusions

This study set out to examine the productive use of aspects of EFL university learners' lexical and phraseological knowledge. Seven indices (five related to lexical sophistication and two related to bigram use) were examined for between-group comparisons and for their correlations with the holistic writing score, testing their ability to predict the quality of EFL writing.

The results showed that the upper proficiency group (S5) performed significantly better than the lower group (S1) with respect to both the frequency and range of content words. This indicates that even at earlier proficiency levels, better proficiency is associated with the production of more sophisticated content words as operationalized by their frequency and contextual distribution in COCA. Previous studies which examined data from more advanced learners, using corpora such as TOEFL and MELAB essays, highlighted similar findings about the development of sophisticated lexical items as learners enhance their language proficiency (e.g., Crossley et al.,, 2010; Kyle et al. 2018; Kyle et al., 2020; McNamara et al., 2010; Monteiro et al. 2018; Salsbury et al., 2011). It is important to note that our higher proficiency group (S5) have just completed their second year at university, which indicates that over the first two years of studying EFL at university, S5 learners managed to enrich their lexical production and to use more sophisticated vocabulary compared to university freshmen (S1).

The significant between-group differences which the range and frequency of content words showed are reflected also in their significant correlations with the writing score. The negative correlation of these two indices with students' scores seems to indicate that the use of more contextually restricted and less frequent content words enhances the sophistication of the production and increases one's score (Kyle & Crossley, 2015, 2016). This also suggests that the range and frequency of content words might be considered important indicators of writing quality.

Unlike content words, neither the frequency nor the range of function words showed significant between-group differences. This might be an expected result since function words are widely used in a variety of texts and contexts. Hence, both their range and frequency distributions are, generally, very similar as reflected in similar S1 and S5 means. The correlational analysis showed that the frequency of function words did not correlate significantly with the writing score. It seems that because learners' productions are very similar with regard to the frequency of function words, raters could not notice any differences that would affect their evaluation. Unlike their frequency, the range of function words showed a significant, though small, correlation with the writing score. This might suggest that although function words might not attract raters' attention in their judgment of essays, the contextual distribution of these words might have some relationship with the perceived quality of an essay.

For the index of LD, the results showed that there wasn't any significant difference between the two groups. This suggests that even after spending two years at university, S5 learners are similar to university freshmen in terms of the variety of the lexical items they produced in their writings. This similarity is reflected also in the absence of any correlations between LD and students' writing scores. It might be possible that S5 students have not developed enough vocabulary to be reflected in more lexically diverse texts compared to university freshmen. A similar result was highlighted by Vidal and Jarvis

(2018) who noted that there was no significant difference between 1st and 3rd year students in their study no matter what measure of LD was used. Although LD is an established indicator of language proficiency (e.g., Crossley & McNamara, 2012; Crossley, Kyle, et al., 2014; Crossley, Salsbury & McNamara, 2014; McNamara et al., 2010), it seems that at early stages of FL learning, learners feel satisfied with the use of the words which first come to their minds in their productions, which leads to the repetitive production of readily accessible tokens. In the current study we analyzed LD in learners' texts irrespective of the differences between content and function words. It might be the case that analyzing specific categories of words separately, such as content words, might lead to differences between the two groups as highlighted by previous research (Jarvis, 2017; Jarvis & Hashimoto, 2021).

For the analysis of learners' productive knowledge of bigrams, the two bigram measures showed significant differences between university freshmen (S1) and S5 students. Previous studies (e.g., Bestgen, 2016a, 2016b; Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014) highlighted that the language of more proficient learners is characterized by the use of higher proportions of rare bigrams (highlighted by MI) and lower proportions of frequent ones (highlighted by the *t*-score). The current study seems to challenge this established claim. Although the more proficient group (S5) performed significantly better than university freshmen (S1) in both categories of bigrams, they produced a larger proportion even in frequent bigrams.

It seems that in non-proficient EFL contexts, comparatively better learners produce larger amounts of bigrams, both of the rare and frequent categories. This might indicate that the productions of a larger proportion of rare bigrams by more proficient learners and a larger proportion of frequent bigrams by less proficient learners, as reported by previous research, might be reliable only in contexts where there is clearly a wider difference in proficiency between learners. This is supported also by our correlational and regression analyses. The proportion of frequent bigrams showed a higher correlation with the writing score and contributed significantly to its prediction, unlike the proportion of MI bigrams. This might suggest that the *t*-score, which mostly highlights high frequency combinations, might be a better indicator of proficiency. It might also be the case that raters of essays produced by lower proficiency learners are more influenced by the number of accurate grammatical combinations. Probably, in less proficient EFL contexts, more attention is paid to grammatical combinations compared to content word combinations. Similar to our results, both Kyle and Crossley (2015) and Yoon (2018) reported that higher proportions of rare combinations may not always be an indicator of better L2 proficiency.

The current study has some pedagogic implications. First, the results suggest that better language proficiency is linked to the use of more sophisticated vocabulary both at the level of word range and frequency even at less proficient levels. Hence, it is important to provide learners with instructional opportunities not only to widen their vocabulary knowledge with more words, but also to focus on the frequency and range distributions of words even at the early stages of second language learning. More practically, classroom materials should be designed in such a way that the contextual distribution of words is taken into consideration, besides the already established importance of frequency. This is also likely to improve the perceived quality of learners' writing as we observed that both the indices of the range and frequency of content words correlated significantly with the writing score. Unlike previous research, LD couldn't be highlighted as a good indicator of

writing quality. It might be the case that essay raters tolerate word repetitions in the writings of less proficient writers.

Of more importance is the role of bigram knowledge in ameliorating the quality of a piece of writing. Both the proportion of rare and frequent bigrams correlated with essay scores. This suggests that producing texts with higher levels of strongly associated combinations gives the raters a positive image about the quality of the text. The ability of frequent bigrams to significantly predict the quality of the writing score might suggest that raters are more interested in noticing more accurate frequent and grammatical combinations in one's writing.

Although this study has highlighted important findings, there are some caveats that need to be addressed in future research. In its analysis of LD, the study has dealt with all the words together, which might have affected the conclusions we have reached in the sense that important between-group differences might have been noticed if content words were analyzed separately. This study has made the conclusion that the production of frequent bigrams, mainly composed of grammatical or function words, contributes significantly to essay quality. We have not, however, examined which function words contribute to essay quality. Future research can deal with this aspect of bigram knowledge. Previous research (Jarvis & Hashimoto, 2021) concluded that the operationalizations of word types differently might yield different LD results. Hence, the conclusions of the current study are valid only as far as operationalizing word types as lemmas is concerned. Future research may examine LD in the writings of Moroccan EFL learners by comparing the results of other operationalizations such as orthographic forms, word families, and flemmas. Although we reached quite an acceptable level of inter-rater reliability in scoring the students' essays, future research could aim at better inter-rater reliability. The conclusions which this study has outlined might apply only to the writing genre which is adopted for this study (expository writing). Further research might compare the production of learners in the target indices using different genres.

Acknowledgments

We would like to thank the two anonymous reviewers for their insightful comments and suggestions. Their feedback has certainly improved the quality of this article.

Correspondence should be addressed to Brahim Ait Hammou.
Email: hammou76@gmail.com

References

Bestgen, Y. (2016a). Using collocational features to improve automated scoring of EFL texts. *Proceedings of the 12th workshop on multiword expressions,* 84-90.

Bestgen, Y. (2016b). Evaluation automatique de textes : Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère. *Traitement automatique des langues, 57*(3), 91-115.

Bestgen, Y. (2019). Évaluation de textes en anglais langue étrangère et séries phraséologiques : comparaison de deux procédures automatiques librement accessibles. *Revue française de linguistique appliquée*, *XXIV*, 81-94. https://doi.org/10.3917/rfla.241.0081

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26,* 28-41. https://doi.org/10.1016/j.jslw.2014.09.004

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* Academic Press.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The moving average type token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94-100. https://doi.org/10.1080/09296171003643098

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The encyclopedia of language and linguistics* (pp. 3168-3171). Oxford University Press.

Cowie, A. P. (Ed.). (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press.

Crossley, S. A., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*, *38*, 46-50. https://doi.org/10.1016/j.asw.2018.06.004

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading. 35*(2). 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Cai, Z., & McNamara, D. (2012). Syntagmatic, paradigmatic, and automatic N-gram approaches to assessing essay quality. *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference, FLAIRS-25* (pp. 214-219).

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*, 573-605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

Crossley, S. A., Salsbury, T., & McNamara D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*(5), 570 -590, https://doi.org/10.1093/applin/amt056

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning: What predicts early lexical production? *Studies in Second Language Acquisition, 35*(4), 727-755. http://doi.org/10.1017/S0272263113000375

Crossley, S. A., Kyle, K., Allen, L., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment, 7*(1). https://escholarship.org/uc/item/06n1v820

Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). *Modelling and assessing vocabulary knowledge*.. Cambridge University Press. doi:10.1017/CBO9780511667268

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics, 19*(4), 443-477. https://doi.org/10.1075/ijcl.19.4.01dur

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 (IRAL), 47*,157-177. https://doi.org/10.1515/iral.2009.007

Ellis, N. C. (2002a). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188. https://doi.org/10.1017/S0272263102002024

Ellis, N. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 297-339. http://doi.org/10.1017/S0272263102002140

Evert, S. (2005). *The statistics of word cooccurrences: Words pairs and collocations* [Unpublished doctoral dissertation]. Universität Stuttgart.

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 1211-1248). Mouton de Gruyter.

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*(2), S397-S408. 10.1044/1058-0360(2013/12-0083)

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research: JSLHR*, *58*(3), 840-852. https://doi.org/10.1044/2015_JSLHR-L-14-0280

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In S*tudies in linguistic analysis* (pp. 1-32). Blackwell.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67*, 155-179. https://doi.org/10.1111/lang.12225

González, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal*, *8*(4), 899-919. https://doi.org/10.1002/tesj.342

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching (IRAL), 52*(3), 229-252. https://doi.org/10.1515/iral-2014-0011

Granger, S. & Bestgen, Y. (2017). Using collgrams to assess L2 phraseological development: A replication study. In P. de Haan, S. van Vuuren & R. de Vries (Eds.), *Language, learners and levels: Progression and variation. corpora and language in use Proceedings 3* (pp. 385-408). Presses Universitaires de Louvain.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*, 403-437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: Further explorations. In *Corpus-linguistic applications* (pp. 197–212). Brill Rodopi. https://doi.org/10.1163/9789042028012_014

Gries, S. T., & Ellis, N. (2015). Statistical measures for usage-based linguistics. *Language Learning, 65*(1), 228-255. https://doi.org/10.1111/lang.12119

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218-238. https://doi.org/10.1016/j.asw.2013.05.002

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. Bazell, J. Catford, M. A. K. Halliday, & R. Robins (Eds.), *In memory of J. R. Firth* (pp. 148-162). Longman.

Howarth, P. A. (1996). *Phraseology in English academic writing.* Max Niemeyer Verlag. https://doi.org/10.1515/9783110937923

Howarth, P. A. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44. https://doi.org/10.1093/applin/19.1.24

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* Routledge. https://doi.org/10.4324/9780203327630

Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., & Möller, J. (2021). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education*, *97*, 103216. https://doi.org/10.1016/j.tate.2020.103216

Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.). *Vocabulary knowledge: Human ratings and automated measures* (pp. 13-44). John Benjamins Publishing. https://doi.org/10.1075/sibil.47.03ch1

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing, 34*(4), 537-553. https://doi.org/10.1177/0265532217710632

Jarvis, S. & Hashimoto, B. J., (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research, 7*(1),163-194. https://doi.org/10.1075/ijlcr.20004.jar

Johnson, W. (1944). Studies in language behavior I: A program of research. *Psychological Monographs, 56*(2), 1-15. https://doi.org/10.1037/h0093508

Jung, Y. J., Crossley, S. A., & McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *Journal of Asia TEFL*, *16*(1), 37-52. https://doi.org/10.18823/asiatefl.2019.16.1.3.37

Kim, M., Crossley, S.A. & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal,102*, 120-141. https://doi.org/10.1111/modl.12447

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/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*, 12-24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030-1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Jarvis, S. (2020). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, *0*(0), 1-17. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Allen, L., Guo, L., & McNamara, D. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment*, *7*(1), 1-16.

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

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Palgrave. https://doi.org/10.1057/9780230511804

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity [Unpublished doctoral dissertation]. The University of Memphis. https://www.proquest.com/docview/305349212

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459-488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, Vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57-86. https://doi.org/10.1177/0741088309351547

Mel'cuk, I. (1995). Phrasemes in language and phraseology in linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schreuder (Eds.), *Idiom: Structural and psychological perspectives* (pp. 167-232). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781315806501

Mel'cuk, I. (1998). Collocations and lexical functions. In A.P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 23-53). Clarendon Press.

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2018). In search of new benchmarks: Using L2 lexical frequency and contextual diversity indices to assess second language writing. *Applied Linguistics*, *41*(2), 280-300. https://doi.org/10.1093/applin/amy056

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32,* 130-149. http://doi.org/10.1017/S0267190512000098

Qian, Y. (2019). Dynamism of collocation in L2 English writing: A bigram-based study. *International Review of Applied Linguistics in Language Teaching,0*(0). https://doi.org/10.1515/iral-2019-0012

Read, J. (2000). *Assessing* v*ocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics, 7*(1), 140-162. https://doi.org/10.1075/arcl.7.06rom

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research 27*(3), 343-360. https://doi.org/10.1177/026765831039585

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan. http://dx.doi.org/10.1057/9780230293977

Sinclair, J. M. (1987). Collocation: A progress report. In R. Steele & T. Threadgold (Eds.), *Language topics* (pp. 319-331). Benjamins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus, 9*(1), 75-106.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209-243. http://doi.org/10.1075/ijcl.8.2.03ste

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. S. Jarvis & M. Daller (Eds.). *Vocabulary knowledge: Human ratings and automated measures* (pp. 79-104). John Benjamins Publishing. https://doi.org/10.1075/sibil.47.05ch3

Vidal, K., & Jarvis, S. (2018). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, *24*(5), 568-587. https://doi.org/10.1177/1362168818817945

Wolk, K., Wolk, A. & Marasek, K. (2017). Unsupervised tool for quantification of progress in L2 English phraseological. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems,* 383-388.

Wray, A. (2000). Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics, 21*(4), 463-489. https://doi.org/10.1093/applin/21.4.463

Yoon, Hyung-Jo. (2018). The development of ESL writing quality and lexical proficiency: Suggestions for assessing writing achievement. *Language Assessment Quarterly, 15*(4), 387-405. https://doi.org/10.1080/15434303.2018.1536756

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Appendix

A list of the most strongly associated rare (MI score) and frequent (t-score) bigrams.

S1			
bigram	**MI *score***	***bigram***	***t-score***
ultimate goal	9.55	of the	1213.61
huge chunks	7.84	on the	786.32
based on	6.71	going to	581.11
driven by	6.30	have been	520.64
grow up	6.16	is a	518.02
have been	5.91	is not	456.88
sacred temple	5.76	they were	382.29
young people	5.59	of them	268.81
hard work	5.44	to have	259.85
tend to	5.16	of their	258.72

S5			
bigram	**MI *score***	***bigram***	***t-score***
dominant role	6.61	it is	957.39
daily life	6.02	to be	746.33
get married	5.91	they are	547.32
long time	5.78	with the	455.05
young people	5.59	you have	420.34
other cultures	5.58	to get	412.39
good job	5.55	there are	377.48
sum up	5.25	have to	350.73
so many	5.19	some of	306.97
supposed to	5.09	the way	298.62