Relationships between Measures of Phraseological Complexity and Writing Quality in a CEFR Assessment Context

Arab Journal of Applied Linguistics
e-ISSN 2490-4198
Vol. 5, No. 1, June 2020, 63-99

http://www.arjals.com/

Lee McCallum, University of Exeter

Abstract

The present study contributes to understandings of the relationships between bigram and collocation complexity and writing quality by analysing a corpus of student placement tests in the UAE. At the heart of the study lies the need to understand the relationships between bigram and collocation diversity and sophistication. In developing such an understanding, the study extends work in this area by examining an underrepresented CEFR grading context. Using correlation analysis and regression modelling, findings indicate that several bigram and collocation measures correlate positively and negatively with essay grades. When built into regression modelling, the 3 predictors of: number of bigram types, Mean MI bigram type, and the number of non-collocation noun + noun bigram types emerge as significant measures that predict grade variation. The implications of these findings for assessment practices in CEFR-using contexts are discussed.

Keywords: *Bigrams, collocation complexity, CEFR assessment, assessment models.*

Introduction

The study of objective linguistic measures of L2 writing quality has a long history spanning the last half-century (Hunt, 1970) with researchers taking manual (e.g. Taguchi, Crawford & Wetzel, 2013) and computational approaches (e.g. Crossley & McNamara, 2012). Operating under Complexity, Accuracy and Fluency (CAF) dimensions, researchers have used these linguistic measures to study how proficiency judgements are made across a range of language assessment contexts. In this respect, several studies have anchored their work in the Common European Framework of Reference (CEFR) (e.g., Garner, Crossley & Kyle, 2018a, 2018b; Granger & Bestgen, 2014; Paquot, 2018, 2019). The primary benefit behind this CEFR work lies in understanding how using linguistic features relates to writing quality and allows students and practitioners to use this knowledge in their respective studies and assessment practices (Hawkins & Filipovic, 2012). This work further influences rater behaviour because exploring relationships between features and writing quality can highlight the extent raters use or deviate from the grading criteria and therefore allow their mental subjectivities to influence their evaluation (Paquot, 2019). Equally, this work can also use human judgements to train automated scoring machines which are increasingly being used to grade writing scripts in large-scale proficiency examinations (Deane & Quinlan, 2010). This influence improves the accuracy of automated systems while also encouraging more writing practice amongst learners because feedback and grading are immediate (Crossley, Defore, Kyle, Dai & McNamara, 2013).

Recently, focus has moved from single word features to word combinations and aspects of collocation and their relationship to writing quality. Under frequency-based definitions, collocation concerns recurring pre-fabricated sequences that writers learn as wholes rather than individual words (Wray, 2002) with Gries (2008) explaining that these units comprise of the co-occurrence of a lexical item and one or more additional linguistic elements. These collocations are often lexical and comprise of units such as adjective + noun (e.g. *young people*), noun + noun (e.g. *drug addict*) and verb + noun (e.g. *generate power*) pairings (Bestgen & Granger, 2014; Granger & Bestgen, 2014). Collocations are deemed to carry a specific meaning and co-occur together more often than chance would allow and are particularly salient in the mental lexicons of native speakers (Durrant & Schmitt, 2009). The importance of collocation in academic writing has been attested in the last two decades and is considered a key component of native language production (Howarth, 1998) and is believed to also highlight mastery of discourse that is typical of a disciplinary community (Li and Schmitt, 2009). However, the mastery of collocation and aspects of word combinations remain an omnipresent challenge for language learners as it is seen as the last hurdle in attaining near-native like proficiency (Prodromou, 2007).

The study of collocations has also been coupled with a study of other aspects of word combinations including the study of automatically extracted sequences of words or 'ngrams' which may include collocations as well as word combinations that perform purely discourse functions in texts (Crossley, Cai & McNamara, 2012). Granger and Bestgen (2014) acknowledge that these ngram extractions are useful for

quickly training automated scoring programmes while the identification of collocations from these ngrams can help investigate pedagogically motivated units that are of special importance to practitioners and language learners. Numerous studies have investigated the relationships between adjacent 2 word ngrams known as 'bigrams' (e.g. '*in the*', '*of the*', '*ozone layer*', '*he used*', '*do justice*') and fine-grained manually identified collocations and writing quality across CEFR-graded EFL contexts. These studies have shown how both aspects have modest to moderate correlations with writing quality (e.g., Bestgen & Granger, 2014; Bestgen, 2017; Garner et al. 2018a, 2018b; Granger & Bestgen, 2014; Paquot, 2018, 2019).

While these studies have contributed to assessment knowledge, this knowledge has yet to be explored in underrepresented CEFR contexts. In this paper, underrepresented CEFR contexts include those that may operate as English as a Lingua Franca (ELF) speaking contexts and therefore differ from Euro-centric contexts that teach English as a Foreign Language (EFL). ELF contexts are those where English operates as a mode of communication between non-native to non-native users of English as well as native to non-native users when these users share no common first language (Jenkins, 2014). In these contexts, meaning negotiation takes place across spoken and written modes (Davies, 2013). A focus on ELF contexts is warranted because in these contexts, language users negotiate meaning by creating their own combinations which serve their needs when they communicate with other language users (Fussell, 2011; Lowenberg, 2002). Prodromou, cited in Prodromou (2007), explains how native speaker English is perhaps inappropriate to

the needs of many learners who use English as a lingua franca. Indeed, previous approaches to the study of bigrams and writing quality have only focused on CEFR-using EFL contexts and it is worth considering how bigram and collocation use in ELF contexts may differ from EFL contexts and what influence this may have on rater behaviour and the ability to award grades to ELF learner essays. In undertaking such an approach, it is worth remembering that because the CEFR does not hold the ultimate standard as being native speaker competency, it should not be taken for granted that raters in ELF contexts are overly sensitive to atypical uses of language because greater emphasis is assumed to be placed on successful communication rather than attaining an unrealistic native speaker standard (Hamid, 2014).

Taking the above issues into account, the current study therefore contributes to the feature-grade literature by exploring the relationships between bigram and collocation complexity and writing quality in the ELF context of the UAE. Then, the study determines the extent this complexity can be used in a regression model to predict grade variation. In doing so, the study aims to shed light on which bigram and collocation features have a relationship with human judgements of writing quality in an underrepresented CEFR context.

Literature Review

Bigram and collocation complexity

Complexity in L2 research has been widely studied since the 1970s (Hunt, 1970). It is generally understood as a writer's ability to use a sophisticated and diverse range of

grammatical structures and vocabulary (Bulté & Housen, 2014). At present, research on bigram and collocation complexity and their relationship to writing quality is in its infancy however Paquot (2018, p.124) is amongst the first researchers to explicitly define this complexity as: "the range of phraseological units that surface in language production and the degree of sophistication of such units". This definition emphasises two sub-constructs of complexity: the diversity of the units produced and the degree of sophistication these units have. The review that follows outlines the previous approaches taken to the study of this diversity and sophistication in CEFR-using EFL contexts. The review concludes by pointing out why such constructs are worth further investigation in other CEFR-using contexts.

Bigram and collocation diversity

Bigram and collocation diversity measures have been studied much less frequently than sophistication. Diversity in single word work has been defined by Read (2000) as the range of different words in a text and in borrowing this definition, diversity has been operationalised by phraseology-focused researchers as bigram and collocation frequency counts which are divided into types and tokens to gauge the extent that writers repeatedly use the same units (Granger & Bestgen, 2014). Types refers to the number of different combinations found in the learner corpus while tokens refers to all the combinations used. The majority of studies have taken the position that frequency is the key criterion for extraction (Bestgen & Granger, 2014). This has resulted in researchers extracting all recurring bigrams and then classifying

them into collocations and other units of interest (Bestgen & Granger, 2014; Granger & Bestgen, 2014; Paquot 2018, 2019).

Granger and Bestgen (2014) are one of the few researchers[1] to explicitly count the frequency of extracted bigrams and investigate their relationship to writing quality. They divided bigram diversity into the number of types and tokens present in the argumentative texts from 3 ICLE (International Corpus of Learner English) sub-corpora (German, Spanish and French L1s) at intermediate and advanced proficiency levels. They found that advanced texts contained more bigrams per 1,000 words than intermediate level texts.

Findings for diverse use of bigrams and collocational units appears to indicate that there is a higher number of bigrams used in higher rated CEFR learner essays in EFL contexts when compared to lower rated essays. The review now turns to consider how these units differ in sophistication.

Sophistication

In operationalising sophistication, researchers again draw on Read (2000, p.200) who outlines single word sophistication as the: "selection of low-frequency words that are appropriate to the topic and style of the writing, rather than just general, everyday vocabulary". Durrant and Schmitt (2009) operationalised this sophistication by using a reference corpus to tap into the extent 2-word premodifier adjective + noun (e.g. '*sweet child*') and noun + noun (e.g. '*global warming*') combinations were beyond

[1] It is worth pointing out the extraction methods differ between Granger and Bestgen (2014) and Paquot (2018) where Granger and Bestgen (2014) extract units of interest via POS tagging only while Paquot (2018, 2019) uses dependency relations extracted by the Stanford Core NLP (2018) to extract units. Given that Paquot's work does not rely on ngram extraction, its discussion is limited throughout this paper.

everyday use. They used the BNC as a reference corpus to allocate a 'Mutual Information' (MI) score to each pairing. This MI score is an association measure which measures the amount of attraction between the individual words in the pairing[2]. The higher the MI score, the stronger the attraction between the words and the more the pairing becomes thought of not only as a collocation but a collocation that consists of words that have a degree of exclusivity to each other (Durrant & Schmitt, 2009). Researchers have calculated the mean MI for each text across proficiency levels. In their corpus of U.S student writing, Bestgen and Granger (2014) found a significant positive correlation with mean MI and overall essay scores and vocabulary scores. The same significant positive correlation was also found in Bestgen's (2017) CEFR-graded FCE (First Certificate in English) and ICLE (International Corpus of Learner English) texts.

MI scores have also been divided into thresholds. Durrant and Schmitt (2009) grouped scores into 7 thresholds that measured the pairings' collocational strength with MI < 3 taken to be non-collocational while pairings scoring ≥ 3 were taken as being collocational in nature. Granger and Bestgen (2014) simplify these groups into 4 threshold strengths: high, medium, low and non-collocational which are presented in Table 1.

[2] Formula details can be found in Durrant & Schmitt (2009).

Table 1

MI-thresholds (Based on Bestgen & Granger 2014; Paquot 2018)

Collocational strength	MI values	Example collocations
Non-collocations	MI values < 3	he_used (MI: 2.06)
Low collocations	MI values ≥3-4.99	used_to (MI:3.84)
Medium collocations	MI values ≥5-6.99	do_justice (MI:5.95)
High collocations	MI values ≥ 7	illegally_parked (MI:9.72)

Durrant and Schmitt (2009) found that adjacent pre-modifying noun pairs that reached medium and high MI thresholds were more used by native speakers and were found to be more used in genre or discipline specific discourse. Granger and Bestgen (2014) found similar when their MI thresholds were applied in that they found higher graded advanced texts used significantly more high MI collocations than intermediate texts while intermediate texts used more low MI collocation types and significantly more non-collocation types. Durrant and Schmitt (2009) further acknowledge the value of the MI score as its score draws out pairings which comprise of low frequency words which share a degree of exclusive attraction to each other. They draw on Clear's (1993) example of '*taste arbiters*' to show that the individual words are low frequency but in analysing their frequency, 25% of their occurrence takes place together. This has important application because it shows that when we meet 'taste', there is a strong possibility that we will also encounter 'arbiters' in its company. This also taps into how word 1 (or word 2) in the pairing gives rise to an expected occurrence of the other word as a partner.

Granger and Bestgen's (2014) study of adjective modifier combinations also reveals that intermediate texts containing significantly more low MI adjective modifiers than advanced texts while significantly more non-collocational units (where MI was < 3)

also appeared in intermediate texts. Granger and Bestgen (2014) also report no significant increases in intermediate texts for use of high, medium and low MI adverbial modifier types however intermediate texts did contain significantly more non-collocational adverbial modifier types than advanced texts.

More recently, Garner (2018a, 2018b) investigated the relationship between CEFR writing grade scores (CEFR A2-B2) and a variety of association measures in a South Korean context. He found that MI scores are amongst the strongest correlates to writing grade scores.

The operationalisation of complexity and innovative word combinations

Researchers have questioned the influence of: (1) absent bigrams from the reference corpus and (2) bigrams that fail to meet a minimum frequency threshold in the reference corpus and so their MI-score is unreliable. Granger and Bestgen (2014) label below threshold units as those appearing ≤ 5 times in the reference corpus (The Contemporary Corpus of American English (COCA) is used in the former and the British National Corpus (BNC) in the latter study). Bestgen and Granger (2014) found COCA absent units negatively correlated with writing quality meaning that as their frequency increased, essay score decreased. A more varied pattern emerges in Bestgen's (2017) study of FCE and ICLE texts whereby a significant negative correlation between writing scores and the proportion of absent units was found for FCE texts while a non-significant positive correlation was found for ICLE texts. Perhaps surprisingly, Granger and Bestgen (2014) found advanced learners used more below threshold combinations from the BNC corpus in total and across

syntactic structures. Granger's work points out an interesting observation in that when both categories were examined they were thought to include erroneous and creative combinations that were markedly different from attested native combinations. Erroneous combinations are described as those which violate grammatical rules (e.g. '*everything are'* as opposed to *'everything is'*) while creative or innovative combinations are combinations which are grammatically possible but are either absent or low frequency occurrences in the reference corpus (e.g. *'ejected students'* as opposed to the more frequent *'suspended students'* or *'expelled students'*) suggesting advanced learners may experiment more with units than intermediate learners who may hold onto memorised choices in their lexicon (Hasselgren 1994).

The pervasiveness of non-collocations, absent and beyond threshold units across writing contexts

Given these findings, the current study views non-collocations, absent and beyond threshold types as units that could potentially offer a window into understanding features of ELF in the UAE. In making this conceptualisation, we can understand their relationship to assessed academic writing that takes place in this context. The current study takes this position in light of many ELF scholars pointing out that these combinations, that are prominently referred to in traditional Learner Corpus Research may, in fact, be investigated to establish whether they can serve as candidates for profiling ELF (Mauranen, 2003).

These findings are relevant to understanding ELF because language in ELF contexts is influenced by learner creation whereby the learning context and learners

themselves are influenced by rich input from non-native and native speakers of English (Lowenberg, 2002). Learners' language production therefore reflects how their context differs from traditional EFL contexts that are more stringently based on standard models of English and adherence to native norms (Bamgbose, 1998). In ELF contexts, learners often produce innovations which are marked differences from the native norms produced by native speakers. Instances of these combinations are evident in Fussell's (2011) study of Gulf English where users of English pluralise uncountable nouns such as: '*homeworks*' and '*furnitures*' and produce specific word pairings: '*road deviations'* instead of *'diversion'*.

The present study therefore seeks to determine if any of the established CEFR EFL relationships between grades and bigrams and/or collocations hold true in an ELF context; and if they do, what can be inferred from them in terms of how raters perceive aspects of ELF language use? The study therefore addresses the following research questions:

1. What is the relationship between bigram complexity measures and writing quality in CEFR graded texts from an ELF context?

2. To what extent do non-collocational, absent and below threshold learner combinations, that may be particularly indicative of ELF use in the UAE, play a role in this relationship?

3. To what extent can these measures predict grade variation in CEFR graded texts?

Methodology

The study focuses on objective bigram measures and follows Granger and Bestgen's (2014) methodology by extracting and analysing crude counts of bigrams as well as detailed counts of bigrams that meet or do not meet collocational status. Research questions one and two are answered by conducting initial Spearman rho's correlations between essay grades and bigram measures. The strongest correlations are then used in regression modelling to answer research question three.

Writing context

The CEPA (Common Educational Proficiency Assessment) - English examination (now rebranded EmSAT, (Ministry of Higher Education and Scientific Research, 2019)) is a university placement exam that assesses grammar, vocabulary, reading and writing. The test acts as a placement test for entry into the UAE's government universities. The test comprised of a 45-minute multiple choice grammar and vocabulary test, a 45-minute multiple choice reading test and a 30-minute writing test (Coombe & Davidson, 2014).

Each university sets its own benchmark for direct entry into its degree programmes with applicants who fall below the benchmark required to complete the university's 1-year preparatory language programme. This programme aims to increase students' proficiency in English through intensive instruction. Students who achieve between an overall grade 4 or 5 (CEFR B1/B2) in the test may be granted direct entry into their degree programmes (Coombe & Davidson, 2014). The CEPA grading rubric only makes sporadic reference to word combinations with individual words

given more focus. Word combinations are only alluded to in grade 1 texts where learners are unable to make 'multi-word messages' while no mention of word combinations is found at grades 2, 3 and 4 while at grades 5 and 6 texts seem more 'register appropriate' with word choice and idiomaticity giving the text fluency. However, overall, the grading rubric has a much stronger focus on individual words, grammar and sentence structure with 'complex sentences used but not always accurately' (See Table 2 in Appendix 1).

On the writing test, students wrote an argumentative essay from a choice of 4 prompts. Topics included the challenges teenagers currently face and favourite time of day. Task requirements specified students should write 150-200 words in their response. While, the examination could be administered as a paper or computer-based test, most tests were computer-based and so these were analysed. Essays were graded by at least two raters. These raters were all experienced teachers in the UAE who had been trained in marking CEPA scripts.

Corpus creation

Texts were taken from a single test administration from locations across the UAE. The Ministry of Higher Education and Scientific Research provided the full test set totalling 1,170 texts. These texts had been rated according to a 1-6 bandscale that was cross-referenced to the CEFR levels A1 to C2 with the highest band 6 equalling CEFR C2 level. As a reliability requirement, grades were checked by two raters with each test administration's essays reaching an inter-rater reliability of 96-98%. In cases of rater disagreement, a third rater is employed to provide a final agreed grade. The

holistic grading rubric required raters to grade according to an overall impression based on task completion, punctuation, spelling and vocabulary use.

The final corpus contained 1,013 texts from grades 2-5 with split grades between these (e.g. 2.5, 3.5). Texts that received grades 0, 1, 1.5, 5.5 and 6 were excluded from analysis as grade 0 texts did not contain enough recognisable English to be analysed and similarly texts at grades 1 and 1.5 contained less than a single sentence in English or simply repeated the test prompt. There were only a handful of texts at grades 5.5 and 6 and so these did not allow inferential conclusions to be drawn given their rarity in the test set.

The corpus is shown in Table 3.

Table 3

Corpus overview

Grade level	Number of texts	Total words	Mean text length (number of words)
2	91	14,283	155.25
2.5	130	23,295	179.19
3	209	43,636	223.14
3.5	207	46,053	223.56
4	195	45,240	232
4.5	114	26,935	236.27
5	67	16,405	244.85
Totals	**1,013**	**215,847**	

By including CEFR levels A2-B2, the study represents a typical test administration. Table 3 indicates most students achieved grades between 2.5 and 4.5 (CEFR levels A2 – B2). For this reason, the corpus maximises representativeness because it reflects actual grade breakdown and shows how many students achieve the threshold for direct entry in a single test administration. Topic selection was fairly even with

prompt 1 selected 251 times, prompt 2: 262 times, prompt 3: 254 times and prompt 4: 246 times across grade levels.

Pre-processing texts

Before extracting bigrams and counting their sub-types, texts were standardised. All titles, headings or sub-headings were removed to avoid inflating phraseological counts. Spelling errors were corrected to American spelling to standardise counts of types. Each corrected text was part-of-speech tagged with CLAWS 7 (UCREL, 2019). While other more modern taggers are in circulation (e.g. the Stanford Part-of-Speech Tagger 2019), CLAWS 7 has a longer history of accurately tagging learner texts (cited as accurately as 97%) (Bestgen & Granger, 2014). Tagging differentiated between grammatical and lexical combinations so pairings such as '*prepare for*' were differentiated from the adjective +noun, noun + noun and verb + noun combinations that the study focuses on.

Extracting and categorizing ngrams

KfNgram (Fletcher 2002-2007) was used to automatically extract all recurrent bigrams (bigrams appearing ≥2) from each text for all grades. *KfNgram* was chosen as it is freely available and has a simple user interface. Bigrams containing numbers and proper nouns were excluded from analysis with a stoplist. Further manual checks ensured these instances were excluded. Numbers were excluded as they were not viewed as modifying structures that the study was interested in and place and personal names were excluded to preserve student anonymity.

Determining complexity measures

Complexity measures for types were operationalised by counting types per text because this allowed variation within each grade to be best understood. A specific focus on types was adhered to as these have been shown to correlate more strongly with grades than tokens in quality-oriented studies (Bestgen & Granger 2014). Additionally, in CEFR studies concerning individual word diversity, type diversity was a stronger predictor of grade scores than token diversity in the Pearson English Academic test (Treffers-Daller, Parslow, & Williams, 2018).

Diversity measures focus on numerical quantitative counts from the learner corpus for example: number of bigrams, number of adjective + noun bigrams, number of noun + noun bigrams and number of verb + noun bigrams.

Final diversity measures are presented in Table 4.

Table 4

Diversity measures (# = number of)

Count types	Measures
Bigram types	# types
Adjective + noun bigram types	# adjective + noun bigram types
Noun + noun bigram types	# noun + noun bigram types
Verb + noun bigram types	# verb + noun bigram types

Sophistication was operationalised using MI scores. Each extracted bigram was allocated an MI-score from COCA. COCA was chosen because learner texts overwhelmingly used American vocabulary as opposed to British or another standard native language variety and COCA was the largest available corpus to identify the possibility of non-native learners using non-standard language (totalling

450 million words). MI-scores were classified using Granger and Bestgen's (2014) collocation strength framework outlined in Table 1.

Sophistication measures included pooled measures derived from the reference corpus: mean MI, mean MI for adjective + noun types, noun + noun types and verb + noun types as well as numerical counts of MI for all types :number of high MI collocation types, medium MI collocation types, low MI collocation types, non-collocation types, absent types and below threshold types. Sophistication measures were also then grouped into MI thresholds based on their syntactic pairings with all sophistication measures shown in Table 5.

Table 5

Sophistication measures (# = number of, ∑=sum)

Mean MI types	Measures	Calculations
Mean MI	Mean MI type	∑ MI for all types ÷ # types
adjective + noun Mean MI	Mean MI adjective + noun types	∑ MI for all adj + noun types/# adj + noun types
noun + noun Mean MI	Mean MI noun + noun types	∑ MI for all noun + noun types/# noun + noun types
verb + noun Mean MI	Mean MI verb + noun types	∑ MI for all verb + noun types/# verb + noun types
High MI collocation types	# High MI collocations	# bigrams with MI score ≥ 7
High MI adjective + noun collocation types	# High MI adjective + noun collocations	# bigrams with MI score ≥ 7
High MI noun + noun collocation types	# High MI noun + noun collocations	# bigrams with MI score ≥ 7
High MI verb + noun collocation types	# High MI verb + noun collocations	# bigrams with MI score ≥ 7
Medium MI collocation types	# Medium MI collocations	# bigrams with MI score: 5-6.99
Medium MI adjective + noun collocation types	# Medium MI adjective + noun collocations	# bigrams with MI score: 5-6.99
Medium MI noun + noun collocation types	# Medium MI noun + noun collocations	# bigrams with MI score: 5-6.99
Medium MI verb + noun collocation types	# Medium MI verb + noun collocations	# bigrams with MI score: 5-6.99
Low MI collocation types	# Low MI collocations	# bigrams with MI score: 4.99-3
Low MI adjective + noun collocation types	# Low MI adjective + noun collocations	# bigrams with MI score: 4.99-3
Low MI noun + noun collocation types	# Low MI noun + noun collocations	# bigrams with MI score: 4.99-3
Low MI verb + noun collocation types	# Low MI verb + noun collocations	# bigrams with MI score: 4.99-3
Non-collocation types	# Non-collocations	# bigrams with MI score: < 3
Non-collocation adjective + noun types	# Non-collocation adjective + nouns	# bigrams with MI score: < 3
Non-collocation noun + noun types	# Non-collocation noun + nouns	# bigrams with MI score: < 3
Non-collocation verb + noun types	# Non-collocation verb + nouns	# bigrams with MI score: < 3
Absent bigram types	# absent types	# types not found in COCA
below threshold bigram types	# below threshold types	# types occurring ≤ 5 times in COCA

In applying Granger and Bestgen's (2014) collocation thresholds, high, medium, low and non-collocations were determined with sample extracted bigrams presented in examples (1) – (4):

(1) **High MI collocations**: *attract tourists*, *bad habits*, *higher levels*, *luxury cars*, *mechanical engineer*, *social media*, *social worker*, *solving problems* and *quit smoking*.

(2) **Medium MI collocations**: *private schools*, *high marks*, *tv show*, *scary movie*, *young adult*, *English teacher*, *family member*, *fashion industry*, and *earn money*.

(3) **Low MI collocations**: *bad things*, *bakery shops*, *play piano*, *huge problem*, *dangerous place*, *high price*, *food store*, *difficult challenges* and *big problem*.

(4) **Non-collocations**: *being famous*, *easy job*, *facing problems*, *good person*, *good salary*, *good time*, *help people*, *new friends*, *rich person*, and *restaurant food*.

Bigrams which occurred ≤ 5 times in COCA were labelled as 'below threshold' and those which did not occur in COCA were labelled 'absent types' meaning bigrams including those in examples (5) and (6) were identified:

(5) **Below threshold types*:*** *smoking shisha*, *putting makeup*, *Turkish movies*, *handmade cooking*, *make BBQ*, *saving humans*, *healthy restaurants* and *caught criminals*.

(6) **Absent bigram types*:*** *unlocal people*, *eating drugs*, *go university*, *made arts*, *extreme important*, *staying calms*, *selling buildings*, *none time* and *play exercise*,

After normality checks, non-normal distribution meant Spearman's rho correlations were run on the measure set with SPSS to answer research questions one and two. The significant correlations were then entered into a regression analysis. To answer

research question three, Tabachnick and Fidell's (2014) cross validation procedure was followed. The data set was split into an 80% training and 20% test set with the former highlighting the strongest candidate measures that may be able to predict grade variation and the latter was then used to verify the model's ability to predict grade variation in an independent data set.

Results and Discussion

Relationships between bigrams and collocations and writing quality

Table 6 shows the significant Spearman rho's correlations between measures of complexity and holistic essay scores. Correlations are organised by their strength and Spearman rho's correlation coefficients are denoted with one asterisk for significance at $p \leq 0.05$ and two asterisks for significance at $p \leq 0.01$ level.

Table 6

Spearman's rho significant correlations with essay grade for training set (# = number of)

Measure	r- value	p-value
Mean MI types	.419**	0.000
# non-collocation types	-.241**	0.000
# medium collocation types	.206**	0.000
# bigram types	-.161**	0.000
# below threshold types	-.129**	0.000
# absent types	-.116**	0.001
# low collocation types	.112*	0.002
Mean MI adjective + noun types	.096*	0.002
# high collocation types	.088**	0.001
# high collocation adjective + noun types	.086**	0.000
# Non-collocation noun + noun types	.085*	0.007
# Low collocation adjective + noun types	.079*	0.003
# Low collocation verb + noun types	.077*	0.003
# Low collocation noun + noun types	.076*	0.003
# adjective + noun bigram types	.069*	0.047
# medium collocation noun +noun types	.066*	0.048

Several complexity measures yield significant correlations with holistic grade scores. However, it should be noted that many of these measures have low r-values (r = <0.10), and therefore their practical value in understanding collocation relationships to writing quality grades may be actually negligible in the minds of raters. However, an important observation can be made with respect to diversity measures in that only bigram types (r_s=-.161) and pooled adjective + noun bigram types (r_s=.069) reach significance. The negative correlation between bigram types and grades suggests that as the bigram types increase, grade scores decrease whereas as the pooled adjective + noun bigram types increase, grade scores also increase.

Sophistication measures yield largely significant positive correlations with holistic scores. The strongest of these correlations is found between Mean MI and holistic scores (r_s=.419). However, fine-grained Mean MI for verb + noun, noun + noun and adjective + noun bigram types do not reach significance.

Combinations that meet collocation status

When the MI thresholds were applied to the combinations, significant relationships that varied in direction emerged. In meeting collocational status, the number of high (r_s=.088), medium (r_s=.208) and low (r_s=.112) collocations all yield significant positive correlations with holistic essay scores. It is noteworthy that there are stronger correlations between medium and low MI collocations than high MI collocations. This observation may be task related as the task relates to real-life experiences where students are expected to argue a position. It may be that a range of more generic, less-exclusive collocations are to be expected and to raters these are deemed more

task appropriate than using high MI collocations which are known to have narrow genre/disciplinary uses (Granger & Bestgen, 2014).

A closer examination of the syntactic patterns reveals that only the number of high MI adjective + noun collocations has a weak significant positive correlation with holistic scores (r_s=.086) while there are only weak significant positive correlations for the number of medium noun + noun collocations (r_s=.066). There are more consistent correlations between holistic scores and low MI syntactic collocations with weak significant positive correlations between holistic scores and low MI adjective + noun collocations (r_s=.079), verb + noun collocations (r_s=.077) and noun + noun collocations (r_s=.076). These correlations highlight several observations regarding ELF learners' uses of collocations and how they appear to be evaluated by raters. The positive correlations between MI thresholds that meet collocational status suggest that collocation use that is attested by presence in a native corpus may be well-received by raters as an increase in their use coincides with an increase in holistic score. There appears to be an indication that syntactic types may also be well received however a lack of correlation between high MI noun + noun and verb + noun collocations and holistic scores may indicate that these patterns are overlooked by raters in favour of medium or low MI collocations which yield stronger correlations to grades. These results partly align with previous mainstream EFL contexts with Granger and Bestgen (2014) also finding that higher proficiency grades used more collocations that achieved high to medium MI threshold levels.

Non-collocation, absent and beyond threshold combinations

Turning to units that fail to meet collocational status, several ELF relevant observations can be made in interpreting how learners in the UAE combine words and how these combinations may be evaluated by raters. First, the number of non-collocation bigram types has a modest significant negative correlation with holistic essay score (r_s=.-.241) indicating that generally, non-attested combinations that do not meet collocational status may not be well-received by raters because an increase in their use occurs as scores decrease. This finding is also consistently found in EFL studies (Bestgen, 2017, Bestgen and Granger 2014, Granger and Bestgen 2014 and Paquot, 2019) strengthening the notion that word combinations that fail to reach collocational status may be perceived negatively by those evaluating writing across EFL and ELF contexts. This observation also sheds light on learner combinations because these combinations occur in both the CEPA corpus and COCA reference corpus but do not have a high enough MI score to meet collocational status.

In examining syntactic types, it is notable that the only non-collocation type that correlates with holistic essay score is the number of non-collocation noun + noun combinations. This is interesting considering the relationships found between collocational syntactic types and may indicate that raters could be particularly sensitive to malformed noun + noun combinations.

Turning to below threshold units, below threshold types yield a weak significant negative correlation with holistic essays scores (r_s=-.129). Along with the finding for non-collocation types, these occurrences suggest that below threshold combinations

with an MI value of < 3 are also viewed less favourably than combinations which are highly frequent and meet collocational status. A similar picture emerges with COCA absent units where these units also have a significant negative correlation with holistic scores (r_s=-.116). Referring back to Granger and Bestgen's (2014) observation that these units consist of erroneous and creative combinations, we can assume that in the UAE, learner innovations may be being evaluated less favourably than combinations which meet collocational status in the native reference corpus.

Explaining grade variation

Those variables yielding significant correlations with grade level were then entered into a training set multiple regression model. Correlations revealed multicollinearity between the number of types and the number of non-collocation types (r-values reached >.70) and so the number of non-collocation types was eliminated from further analysis as its correlation with other type variables also approached r =.70 (following multicollinearity advice in Crossley & McNamara, 2012).

Table 7 shows that the training set correlations produce a significant model with the mean MI for all types, the number of non-collocation noun + noun types and the number of bigram types appearing as significant predictors.

Table 7
Training set model summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Sig
1	.456	.208	.205	.7958	.000

The model explained 20.8% of the variation in grade scores in the training set. The coefficients of the training model were checked for variance inflation factors (VIF)

and tolerance values to assess model reliability and the presence of multicollinearity. The training model did not suffer from multicollinearity as the VIF values were less than 2.50 and tolerance levels were below 0.4 which Jeon (2015) supports as being a valid model that shows no signs of multicollinearity.

Using the beta weights and the constant from the training model in Table 8, the reliability of the model was tested on the test set data. The coefficient values were used to create predicted grade scores which were then compared to the actual CEPA grade scores.

Table 8

Training set model description (B = unstandardized Beta weights, B = standardised Beta weights, SE = standard error). Estimated constant term is 2.399.

Entry	**Variable**	***B***	**B**	***SE***	***t-values***	***Sig***	***Tolerance levels***	***VIF***
1	# bigram types	-.001	-.026	.001	-.834	.000	.987	1.013
2	# non-collocation noun + noun bigram types	-.072	-.104	.021	-3.390	.001	.992	1.008
3	Mean MI type	.556	.437	.039	14.225	.000	.980	1.021

The resulting correlation analysis yielded a moderate correlation between the predicted scores and the actual scores (r=.451). The test set also yielded a significant model which explained 15% of the variation in CEPA grade scores. When the training set Beta weights and constant from the training set were applied to the full data set, a significant model was also found which explained 18% of the variation in CEPA grade scores. The application to the full data set allows the model to be tested on the whole CEPA corpus. The similarity in variance from the training-test-full set

models allows more confidence that the model is reliable across the CEPA essay set (Crossley & McNamara, 2012).

Conclusion

This study aimed to examine relationships between bigram and collocation complexity and writing quality. The study sought to advance knowledge in this area by examining an underrepresented ELF assessment context. The study highlighted several relationships between these constructs. The key relationships centred on the use of sophisticated collocation use determined by examining the MI scores between collocations. Overall, findings indicate that bigrams that meet collocational status seem to positively correlate with grades whereas bigrams that do not meet collocational status seem to negatively correlate with grades in this ELF context. This follows traditional EFL contexts whereby assessment adheres to traditional native models of accuracy and norm standards. This picture was strengthened by examining absent and below threshold units which were also negatively correlated with holistic grade scores. These units contain innovative bigrams which suggests these units may not be welcomed by raters who prefer or expect established collocations to be used. These two findings alone have important implications for the debate on use of language models or norms in assessment contexts. These findings indicate that in the UAE, there appears to be a negative relationship between learner innovations and grade scores and questions what learner innovations, if any, are permissible in this context where negotiation of communication takes place between language users of different backgrounds and competency levels. In this respect, a

limitation of this quantitatively-heavy work, is the lack of direct focus on how raters actually view and experience reading instances of learner errors/innovations and how raters' own assessment background and first language influence this.

The study also aimed to predict grade variation across the CEPA essay set. This resulted in the large-grained measures of: number of bigram types, number of non-collocational noun + noun bigram types and Mean MI being able to predict holistic essay scores. These three measures have important findings for automated scoring systems as well as understanding rater behaviour. First, these measures are in, comparison to fine-grained syntactic pairings, easily computed and so can be useful for taking initial steps towards creating an automated grading system for the writing exam. Second, the absence of syntactic pairings from the prediction may mean that the relationships identified in research question one are not strong enough to accurately ascertain grades and therefore bring into question their perceived value by raters. The small correlations obtained also raise the possibility that other linguistic features do indeed have stronger relationships to grades in this context (as seems to be the case indicated in the CEPA rubric). The failure to include other linguistic features of interest from the grading rubric (e.g., features of single word choice, grammatical structures and accuracy features) is a potential limitation of this study that future quantitative work could explore further. Introducing these features into such work is likely to also clarify further the true importance of bigrams and collocations when they are examined with other linguistic features that have been shown predict grade scores in other contexts and studies (e.g., Paquot, 2019).

Equally, it is important to reflect on the fact that the study does not factor in contextual variables into its modelling work. The reliance on monofactorial linear regression modelling in such quantitative grade studies has been questioned in the second language literature (e.g. see Alexopoulou, Michel, Mukarami, & Meurers, 2017) and future grade-feature work could be attempted by using multifactorial methods that take into account factors such as rater experience or rater background and test taker characteristics (e.g. language learning experience) and how these are likely to have an influence on grade allocations.

A final limitation that future research should develop is a closer qualitative examination of the types of non-collocation, absent and beyond threshold units that characterise these categories. A further qualitative analysis would offer a more fine-grained description of word combinations in this ELF context as has been advocated by Seidlhofer (2009). This kind of focus is very much needed if assessment criteria are to align more closely with local academic uses of language in the UAE and the wider MENA region.

Acknowledgements

The Ministry of Higher Education and Scientific Research in the UAE provided the essay set that this study is based on.

References

Alexopoulou , T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*: S1, 180-208.

Bamgbose, A. (1998). Torn between the norms: Innovations in world Englishes, *World Englishes*, *17*(1), 1-14.

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.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65-78.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65.

Clear, J. (1993). Tools for the study of collocation. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and technology: in honour of John Sinclair* (pp. 271-292). Amsterdam: John Benjamins.

Coombe, C., & Davidson, P. (2014). Common Educational Proficiency Assessment (CEPA) in English. *Language Testing*, *31*(2), 269-276.

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.

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

Crossley, S. A., Defore, C., Kyle, K, Dai, J., & McNamara, D. S. (2013). Paragraph specific N-Gram approaches to automatically assessing essay quality. In S.K. D'Mello., R. A, Calvo & A. Olney (Eds,). *Proceedings of the 6th Educational Data Mining (EDM) Conference.* (pp. 216-220). Heidelberg, Berlin, Germany: Springer.

Davies, A. (2013). *Native speakers and native users: Loss and gain*. Cambridge: Cambridge University Press.

Deane, P., & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research*, *2*(2), 151-177.

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

Fletcher, W.H. (2002-2007). *KfNgram*. Annapolis, MD: USNA.

Fussell, B. (2011). The local flavour of English in the Gulf. *English Today*, *27*(4), 26-32.

Garner, J., Crossley, S., & Kyle, K. (2018a). Beginning and intermediate L2 writer's use of ngrams: An association measures study. *International Review of Applied Linguistics*, Ahead of print. DOI: https://doi.org/10.1515/iral-2017-0089

Garner, J., Crossley, S., & Kyle, K. (2018b). Ngrams and L2 writing proficiency. System, 1-37. DOI: 10.1016/j.system.2018.12.001

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*, *52*(3), 229-252.

Gries, S.T. (2008). Phraseology and linguistic theory: A brief survey. In. S. Granger and F. Meunier (Eds,). *Phraseology: An interdisciplinary perspective* (pp. 3–25). Amsterdam: John Benjamins.

Hamid, O.M. (2014). World Englishes in international proficiency tests. *World Englishes, 33*(2), 263-277.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners. A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237-258.

Hawkins, J.A., & Filipovic, L. (2012). *Criterial features in L2 English: Specifying the Reference Levels of the Common European Framework*. Cambridge: Cambridge University Press.

Hunt, K.W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly*, *4*(3), 195-202.

Jenkins, J. (2014). *English as a lingua franca in the international university: The politics of academic English language policy*. London: Routledge.

Jeon, E.H. (2015). Multiple linear regression. In L. Plonsky (Ed.). *Advancing quantitative methods in second language research* (pp.130-158). New York & London: Routledge.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A

longitudinal case study. *Journal of Second Language Writing*, *18*, 85-102.

Lowenberg, P. (2002). Assessing English proficiency in the expanding circle. *World Englishes*, *21*(3), 431-435.

Mauranen, A. (2003). The corpus of English as a Lingua Franca in Academic Settings. *TESOL Quarterly*, *37*(3), 513-527.

Ministry of Higher Education and Scientific Research. (2019). EmSAT Achieve test details. Available at: http://emsat.moe.gov.ae/emsat/EmSAT_achieve_en.aspx. Last accessed: 28/03/2019.

Ministry of Higher Education and Scientific Research (2019). CEPA-English Public Test Specifications. Accessed at: http://ws1.mohesr.ae/cepa/Files/Public_CEPA_English_Specifications.pdf. Last accessed: 28/03/2019.

Paquot, M. (2018). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121-145.

Paquot M. (2019). Phraseological competence: a useful toolbox to delimitate CEFR levels in higher education? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, 15(1), 29-43.

Prodromou, L. (2007). *English as a Lingua Franca: A corpus-based analysis*. London: Continuum.

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

Seidlhofer, B. (2009). Common ground and different realities: World Englishes and English as a lingua franca. *World Englishes*, *28*(2), 236-245.

Stanford Log-linear Part-of-Speech Tagger. (2019). Stanford POS. Available at: https://nlp.stanford.edu/software/tagger.shtml. Last accessed: 28/03/2019.

Tabachnick, B.G., & Fidell, L.S. (2014). *Using multivariate statistics* (6th edition). Harlow, UK: Pearson Education Limited.

Taguchi, N., Crawford, W., & Wetzel, D.Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly*, *47*(2), 420-430.

Treffers-Daller, J.T., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39* (3), 302- 327.

UCREL. (2019). CLAWS tagger. Available at: http://ucrel.lancs.ac.uk/claws/. Last accessed: 28/03/2019.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

APPENDIX 1: CEPA GRADING RUBRIC

Table 2: *CEPA grading rubric*. Ministry of Education (2014). CEPA – English Public Test Specifications.

Band	Description	CEFR level
0	• Any of the following: - No sample available. - Whole text appears to be copied, scanned or memorized. - Arabic, random typing or illegible handwriting (PBT only)	Less than CEFR A1
1	• Message is frequently unclear. • Essentially unable to make sentences or multi-word messages. • Vocabulary is extremely limited. • Spelling errors are frequent, even in common words, except when copied. • May confuse upper and lower case letters. • Punctuation is mainly missing or inaccurate. • Text may be so short that it is difficult to assess meaningfully.	CEFR A1
2	• Can convey only the simplest ideas. • Attempts to produce short sentences and phrases independently, but with little control of sentence structure. • Vocabulary is limited to common words. • Can spell a few common words accurately. • Some evidence of punctuation, but usually inaccurate.	CEFR A1-A2
	• Meaning is clear in short, straightforward communications but becomes unclear if content is longer or	

3	more complex. • Some attempt to organize ideas, but little evidence of cohesive devices. • Produces simple sentences with some awareness of, but limited control of, basic sentence structure. • Vocabulary is related to the topic but limited to the simplest range. • Spelling of familiar words is generally accurate, but unfamiliar words may be unrecognizable. • Uses capital letters and full stops most of the time.	CEFR A2- B1
4	• Meaning is generally clear, but can become unclear in complex communication. • Simple cohesive devices used appropriately. • Can construct simple sentences but basic errors (e.g. subject verb agreement and tense) still occur. • Attempts complex sentences. • Range of vocabulary becomes wider, but may be inappropriate. • Text is stilted. • Spelling errors can intrude, though words are mainly recognizable. • Uses capital letters and full stops almost without error; commas and apostrophes missing or misused.	CEFR B1-B2
	• Meaning is generally clear and unambiguous. • Main and subsidiary points are generally well organized. • A range of cohesive devices is used, though not always	

5	appropriately. • Simple sentences are generally correct; some complex sentences may be used, but not always accurately. • Basic errors may still occur. • Vocabulary is generally appropriate for the topic. • Appropriate choice of words, idioms and register occasionally gives a sense of fluency. • Spelling errors can intrude, but do not impair meaning. • Punctuation is used appropriately, with only occasional errors.	CEFR B1-B2
6	• Meaning is clear and unambiguous throughout. • Main and subsidiary points are well organized. • A range of cohesive devices is used appropriately. • Generally accurate use of simple and complex sentences. Errors rarely impede understanding. • Vocabulary choice is generally adequate, but may be inadequate to express a wide range of ideas with precision. • Most of the time, appropriate choice of words, idioms and register gives the text a feeling of fluency. • Occasional errors in spelling may occur. • Punctuation is used appropriately.	CEFR B2 - C1+