James Garner*, Scott Crossley and Kristopher Kyle

Beginning and intermediate L2 writer's use of N-grams: an association measures study

https://doi.org/10.1515/iral-2017-0089

Abstract: Acommon approach to analyzing phraseological knowledge in first language (L1) and second language (L2) learners is to employ raw frequency data. Several studies have also analyzed n-gram use on the basis of statistical association scores. Results from n-gram studies have found significant differences between L1 and L2 writers and between intermediate and advanced L2 writers in terms of their bigram use. The current study expands on this research by investigating the connection between bigram and trigram association measures and human judgments of L2 writing quality. Using multiple statistical association indices, it examines bigram and trigram use by beginner and intermediate L1 Korean learners of English in English placement test essays. Results of a logistic regression indicated that intermediate writers employed a greater number of strongly associated academic bigrams and spoken trigrams. These findings have important implications for understanding lexical development in L2 writers and notions of writing proficiency.

Keywords: Learner corpus research, association measures, N-gram analysis, L2 writing proficiency

Writing skills are one of the most important, yet most challenging, abilities to develop for second language (L2) learners (Bialystok 1978; Crossley 2013; Nunan 1989; White 1981). This is especially true for L2 writers in academic contexts who are often required to demonstrate their knowledge of course content through written exams or essays. In addition, learners are often required to take part in standardized writing assessments for admissions, placement, and advancement purposes. As a result, a great deal of research has focused on the relationship between linguistic features in L2 essays and proficiency scores assigned by expert human raters (Crossley and McNamara 2012).

***Corresponding author: James Garner,** Applied Linguistics and ESL, Georgia State University, 15th Floor, 25 Park Place, Atlanta, Georgia, USA, E-mail: jgarner17@gsu.edu
Scott Crossley, Applied Linguistics and ESL, Georgia State University, 15th Floor, 25 Park Place, Atlanta, Georgia, USA
Kristopher Kyle, Second Language Studies, University of Hawaii at Manoa, Honolulu, Hawaii, USA

These linguistic features have generally been either syntactic, cohesion (both local and global), or lexical. Regarding syntactic features, research has shown that L2 texts receiving higher ratings contain a greater number of complex clauses, more subordination, and more passive structure (Crossley and McNamara 2014; Friginal et al. 2014; Grant and Ginther 2000). Findings from studies investigating the connection between the use of local and global cohesion have produced contradictory results. On the one hand, studies have shown that increased use of connectives (Jin 2001), pronouns (Reid 1992), semantic overlap (Crossley et al. 2014), and function word overlap between adjacent sentences and paragraphs (Crossley et al. 2016) leads to higher essay ratings. Conversely, recent studies have reported that higher scoring essays employ less aspect repetition and content word overlap and fewer conditional connectives (Crossley and McNamara 2012; Guo et al. 2013). Lastly, studies of lexical features and L2 writing proficiency have revealed that higher scoring essays contain less frequent, less meaningful, and less familiar words and are more lexically diverse (Crossley et al. 2013; Crossley and McNamara 2012; Crossley et al. 2014; Grant and Ginther 2000; Jarvis 2002; Laufer and Nation 1995).

Most research on lexical sophistication and L2 writing has tended to focus on measures of single word use (Crossley et al. 2012). However, recent research has begun to investigate the connection between L2 writing proficiency and multi-word sequence use (Kyle and Crossley 2015). These multi-word sequences include items such as collocations (i. e. lexically restricted combinations), idioms (i. e. semantically non-compositional phrases), and lexical bundles (i. e. frequently recurring word sequences that fulfill pragmatic or discourse functions). Corpus-based research has revealed the ubiquity of these different multi-word sequences across a range of spoken and written genres (Erman & Warren, 2000; Römer 2009), while psycholinguistic research has provided evidence that knowledge of multi-word sequences can give language users significant processing advantages (Ellis 2012; Siyanova-Chantura and Martinez 2015). L2 writing studies also indicate that production of multi-word sequences is an important indicator of success. For instance, research has shown that, despite difficulties in producing native-like multi-word sequences (Ädel and Erman 2012; DeCock et al. 1998; Granger 1998; Ishikawa 2009; Nesselhauf 2003, 2005), more proficient L2 writers produce a greater range of multi-word sequences common to native speech and writing and produce them more frequently than lower proficiency writers (Ohlrogge 2009; Kyle and Crossley 2015; Vidakovic and Barker 2010).

In most cases, research into L2 writers' multi-word sequences use has followed one of two approaches. In the first approach, often referred to as the

phraseological approach (Granger and Paquot 2008), multi-word sequences are classified on the basis of semantic characteristics (e. g. transparency of meaning and substitutability of constituent words). This approach differentiates between free combinations (*pay money*), restricted collocations (*pay attention*), and idioms (*pay lip service*). The second approach towards analyzing L2 phraseology identifies n-grams, or recurrent sequences of contiguous words, on the basis of whether or not they recur above a certain frequency threshold (Granger and Paquot 2008).

There is a third, lesser used approach that is gaining momentum and that we use in the current study. This approach identifies and analyses learner multi-word sequence production using statistical measures that attempt to quantify the association between contiguous words (Bestgen and Granger 2014; Durrant and Schmitt 2009; Granger and Bestgen 2014). In the current study, we examine the association strength of bigrams (two word sequences) and trigrams (three word sequences) in texts written by beginner and intermediate L2 writers. A large corpus of student essays categorized into three proficiency levels by human raters was analyzed using an automatic text analysis program (TAALES, Kyle and Crossley 2015) that measures bigram and trigram association strength. TAALES includes a wide range of bigram and trigram association strength measures, including many previously not utilized in L2 phraseological studies. The main goal of the current study is to investigate whether or not these bigram and trigram association strength measures can be used to predict human judgments of writing quality. In doing so, this study attempts to better understand the connection between L2 multi-word sequence use and writing proficiency as well as provide alternative n-gram association measures to be used in future n-gram analyses.

1 Lexical sophistication and L2 writing proficiency

Lexical sophistication is often defined using a dual trait model that includes breadth and depth of vocabulary knowledge (Meara, 2005). Breadth refers to how many words a speaker knows and is generally measured using lexical diversity or frequency measures. Depth refers to how well a word is known and can be assessed with lexical properties such as polysemy (i. e., the number of word senses a learner know) or hypernymy (the specificity of words known). Word association strength measures would fit within the criterion of depth of knowledge because the strength of associations between words would indicate greater word knowledge (Read, 2000, 2004). Recent research from corpus-based

and psycholinguistic research highlights links between depth of knowledge and strength of associations. From a corpus-linguistic perspective, research has shown that multi-word sequences form the basis of a large proportion of utterances in English (Erman & Warren, 2000; Römer 2009) and that the types of multi-word sequences vary among spoken and written registers and discourse communities (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Hyland 2008). From a psycholinguistic perspective, studies have shown how knowledge of multi-word sequences can give language users significant processing advantages. Specifically, it has been demonstrated that multi-word sequences are processed quicker in receptive and productive language tasks (Ellis 2012; Siyanova-Chantura and Martinez 2015).

Both breadth of depth of lexical knowledge are important components to understanding L2 writing proficiency. L2 writing research has consistently demonstrated a strong connection between measures of lexical sophistication and human judgments of L2 lexical proficiency and overall writing proficiency. For instance, findings from Crossley, Salsbury, McNamara, and Jarvis (2011) and Kyle and Crossley (2015) have demonstrated that more lexically proficient L2 texts contained a more diverse range of words, words that appear in fewer contexts, and less frequent, meaningful, familiar, and specific words. In regards to overall writing proficiency, Crossley and McNamara (2012) found that higher scoring L2 essays contained a more diverse range of words, less frequent content words, and less familiar and meaningful words. Similar results were found by Crossley et al. (2014), who found that word frequency, meaningfulness, and familiarity were predictive of written responses to the TOEFL-iBT. In the latter two studies studies, lexical indices were the most numerous and most predictive features of human judgments of writing quality compared to syntactic or cohesion measures.

Research on the use of multi-word sequences in L2 writing has revealed that, similar to single-word use, more proficient writers employ a greater range of idioms, collocations, and n-grams. This research has also shown that more proficient L2 writers produce these types of multi-word sequences more frequently than less proficient writers. For instance, Ohlrogge (2009) found that intermediate writers scoring higher on standardized writing assessments employed idioms and collocations more frequently while relying less on multi-word sequences provided in the writing prompt. Similar to studies of single word use, quantitative analyses of n-grams in L2 writing have shown that higher proficiency L2 writers use a greater range of 4-grams more frequently (Vidakovic and Barker 2010). Texts judged to display a high degree of lexical proficiency have also been found to contain a greater proportion of bigrams found in large native-speaker reference corpora and, in contrast to results for

individual words, more frequent trigrams (Kyle and Crossley 2015, 2016). Lastly, research has shown that more proficient writers employ different kinds of n-grams compared to their intermediate counterparts. Chen and Baker (2014) found that texts written by advanced writers contained more n-grams characteristic of academic discourse (e. g. noun-phrase and prepositional phrase n-grams), while intermediate texts contained more n-grams characteristic of spoken discourse (e. g. verb-based n-grams). However, in each of these studies multi-word sequences were identified using either native speaking judges (Ohlrogge 2009) or raw frequency counts of contiguous words (Chen and Baker 2014; Kyle and Crossley 2015; Vidakovic and Barker 2010). Both approaches have limitations. The use of native speaker judges for identifying multi-word sequences can cause some sequences that are not immediate salient to native speakers to be overlooked. The use of raw frequency counts to identify multi-word sequences can cause some highly frequent word combinations to be identified as multi-word sequences simply because their component words are also highly frequent (Evert 2009). For instance, the bigram *at the* is a common bi-gram in English, but most fluent language users would not recognize it as a multi-word sequence, mainly because it doesn't have a strong semantic component.

2 N-gram association measures

To address these limitations, researchers have begun to identify and analyze multi-word sequences using statistical association measures that identify sequences based on the probability of their occurrence in language use. These probabilities compare the actual frequency of the n-gram in a corpus with its expected frequency given the frequency of its component words (Evert 2009). Two of the most commonly used statistical association measures are pointwise Mutual Information (MI) and t-score. These association measures differ in two key ways. Firstly, MI is an effect size measure, while t-score is a measure of significance (Groom 2009). In other words, MI indicates the strength of attraction between two words in a bigram, while t-score measures the certainty with which two words can be claimed to form a bigram. Secondly, MI and t-score differ in the kinds of word combinations they highlight. MI has a tendency to highlight bigrams that consist of low-frequency words (e. g. *densely populated*), while t-score highlights bigrams that consist of high-frequency words (e. g. *for example*). Psycholinguistic research has shown that these statistical association measures more accurately reflect the phraseological knowledge of native English

speakers. In a series of receptive and productive language tasks, Ellis and Simpson-Vlach (2009) found that MI outperformed raw frequency in predicting native speaker performance. Specifically, it was shown that sequences with higher MI scores were recognized and comprehended quicker and produced more quickly when read aloud. High MI sequences also had stronger priming effects in pronunciation. In contrast, the raw frequency of the target multi-word sequences had no significant effects in any of the tasks.

Some researchers have begun to use these association scores to examine L2 writing proficiency. For instance, using MI and t-scores, Durrant and Schmitt (2009) examined the use of premodifier-noun (i. e. noun-noun and adjective-noun) bigrams in the writing of L1 and L2 writers. They reported that L2 writers overused a smaller range of bigrams with higher t-scores and underused bigrams with higher MI scores, indicating that L2 writers tend to over rely on bigrams consisting of high-frequency words and under utilize bigrams composed of low-frequency words. In two similar studies, Granger and Bestgen (2014) and Bestgen and Granger (2014) examined the use of statistically associated bigrams across proficiency levels and longitudinally respectively. Granger and Bestgen (2014) found that intermediate writers employed more high t-score bigrams, while the advanced writers made greater use of bigrams with high MI scores. In addition, intermediate writers produced more non-collocational adjective-noun and adverb-adjective bigrams than advanced writers. In the second study, Bestgen and Granger (2014) investigated bigram production in essays written by university-level ESL writers three times over the course of a semester. The longitudinal analysis revealed a significant decrease in mean t-score between the first essay to the third essay in terms of both bigram types and tokens. In terms of writing proficiency, the authors found strong positive correlations between mean MI scores (type and token) and all three categories of essay quality.

3 Limitations of MI and t-scores

While the use of MI and t-scores as association measures have provided insight into the productive phraseological knowledge of L2 writers, these measures are not without their limitations. First, as previously stated, both measures favor one kind of word combination over another. MI tends to highlight combinations composed of low-frequency words, while t-score highlights those composed of high-frequency words. Second, the t-score statistic assumes a normal distribution of data. This makes it problematic for use as an association measure because natural language data is rarely normally distributed (Evert 2004;

Stefanowitsch and Gries 2003). Last, both MI and t-score are symmetrical association measures and, as such, do not account for the directionality of word combinations (Gries and Ellis 2015). For example, the bigrams *of course* and *course of* would receive the same MI and t-scores, even though most speakers would recognize them as two different bigrams with different strengths of association.

There are several possible solutions to address these issues. To counter problems with MI and its bias towards bigrams with low-frequency words, MI^2 can be used as an alternative (Evert 2009). In comparison to MI, MI^2 increases the influence of the observed n-gram frequency as compared to expected frequency, thus counterbalancing its low-frequency bias. The issue regarding t-score can be addressed through the use of collexeme analysis (Stefanowitsch and Gries 2003). Collexeme analysis calculates the bidirectional strength of association between lexemes using the negative log of the Fisher-Yates exact test (Fisher 1934; Yates 1934). As a result, it does not have the same distributional assumptions and is more appropriate for natural language data. As such, this test can provide a more accurate measure of statistical word association. To account for the directionality of n-grams, Gries (2013) suggests using ΔP, a contingency measure from associative learning research and first introduced to the language acquisition research by Ellis (2006). ΔP calculates the probability of an outcome given a cue minus the outcome occurring without the cue present. In the case of bigrams, ΔP calculates the probability of one word in the bigram given the occurrence of the other word in the bigram (e. g. the probability of *course* given the occurrence of *of*). According to Gries (2013), this can tease apart different associations between words in an n-gram that might be hidden with bidirectional association measures.

4 The current study

Previous research has revealed significant differences in the use of statistically associated bigrams between L1 and L2 writers, between L2 writers across proficiency levels, and as a function of time studying English. However, several aspects regarding the association strength of n-grams employed by L2 writers still require investigation. First, while previous association measure studies have focused on intermediate and advanced L2 writers, the use of n-grams by beginning L2 writers remains under investigated. Research into n-gram use by beginning writers would provide a fuller understanding of L2 productive phraseological knowledge and how it differs across proficiency

levels. Second, previous studies of n-gram association strength in L2 writing have been restricted to the analysis of bigrams and have predominantly relied on just two measures of n-gram association, MI and t-score. Lastly, while several studies have investigated how n-gram frequency and proportion indices can be used to predict human judgments of L1 and L2 writing quality (Crossley et al. 2012, 2014; Kyle and Crossley 2015, 2016), the extent to which n-gram association measure indices can predict human judgments of L2 writing quality has so far received little attention. Such information regarding a wider range of bigram and trigram association measures and their relationship to human judgments of L2 writing proficiency would be important for several reasons. First, it could identify the relationship between n-gram production and human judgments of writing proficiency. This information would provide further support for focusing on these items in L2 pedagogy. In addition, the inclusion of a range of association measures allows for a possible comparison of their predictive power, providing justification for including them in future studies of L2 phraseological knowledge.

To that end, the current study investigates links between statistical association measures for both bigram and trigrams and human judgments of L2 writing quality. A large corpus of essays written by Korean EFL students during an English placement test was analyzed using TAALES, which automatically calculates a range of bigram and trigram association measures. The essays in the current study were assigned to one of three proficiency levels (high-beginner, low-intermediate, high-intermediate) by human raters according to the Central European Framework of Reference for Languages (CEFR). Bigram and trigram association indices that showed significant differences between the three proficiency levels were used in a Logistic Regression to predict human judgments of proficiency level. The research questions of interest for this study are:

1. Are n-gram association measures predictive of human judgments of writing quality and, if so, which measures are most predictive of human judgments of writing quality?

5 Methods

5.1 Corpus

The current study makes use of a sample of texts from the Yonsei English Learner Corpus (YELC; Rhee and Jung 2014). This corpus comprises 6,572 written responses to the Yonsei English Placement Test (YEPT) written by 3,286 Korean

EFL students and totals just over 1 million words. This computer-based test is administered to Korean high school graduates admitted to Yonsei University in Seoul, South Korea, and includes both a spoken and written section. The written portion of the test asks students to respond to three different prompts. Prompt 1 requires students to create sentences based on words presented to them on screen. Prompt 2 asks students to write about a topic familiar to them (e. g. What was your favorite extracurricular activity in high school?) with a suggested word limit of 100 words. Prompt 3 requires students to write about a prompt that is more academic in nature (e. g. Why should people receive a college education? State your opinion) and has a suggested word limit of 300 words (although students were allowed to go over). These texts were assigned a grade by trained native English speaking raters at Yonsei University according to the Common European Framework of Reference for Languages (CEFR). No information about inter-rater reliability among the raters is available. The corpus is divided into nine ordered proficiency levels ranging from A1 to C2 (including levels A1+, B1+, and B2+). Only responses to Prompt 2 (responding to a familiar topic) and Prompt 3 (responding to an academic prompt) are represented in YELC. For the current study, a random sample of 450 Prompt 3 essays were chosen from three proficiency levels in the YELC: high-beginner (A1), low-intermediate (B1), and high-intermediate (B2) based on sample size considerations (i. e., these three levels had large representative sample sizes). Prompt 2 essays were not included in the dataset due to their small text size, with most essays containing less than 100 words. Previous studies have shown that texts shorter than 100 words can cause issues for automatic text analysis tools (Crossley and McNamara 2013). The dataset for this study totaled 1,350 essays and comprised 320,404 words. More information about the essays is presented in Table 1.

Table 1: Descriptive statistics for data.

	A2	B1	B2
Texts	450	450	450
Words	87,933	103,001	129,470
Words/Text	195.41	228.89	287.71

5.2 Text analysis

In order to investigate student bigram and trigram use in their essays, the current study makes use of the Tool for the Automatic Analysis of Lexical

Sophistication 2.0 (TAALES; Kyle and Crossley 2015). TAALES is an automatic text analysis program that reports on indices related to lexical sophistication such as word frequency, range (i. e. in how many texts a word or n-gram occurs in the reference corpus), psycholinguistic word information (i. e. word concreteness, word imageability), and academic language for individual words and n-grams. TAALES 2.0 reports on a number of indices derived from the Corpus of Contemporary American English (COCA; Davies 2009) as well as the BNC as reference corpora. It also includes indices based on the five different subsections of COCA (e. g. academic, fiction, magazines, news, and spoken) rather than the entire corpus altogether.

For the current study, the focal indices are association measures for bigrams and trigrams based on the spoken and academic subsections of COCA. TAALES calculates five association measures for each bigram and trigram found in a text that is also found in the reference corpora. These five measures are Mutual Information (MI), Mutual Information Squared (MI^2), t-score, ΔP, and collexeme score. For each measure, an average association score is then calculated for the text using the scores for all bigrams and trigrams receiving an association score in the text. As previously discussed, MI, MI^2, and t-score calculate the probability of co-occurrence for two words regardless of order, while ΔP calculates the probability of the second word in a bigram given the occurrence of the first word in it. The collexeme association measure (Gries, Hampe, & Shönfeld, 2005) calculates the bidirectional strength of association between lexemes using an approximate score that is highly correlated with actual collostructional strength scores calculated by the negative log of the Fisher-Yates exact test (Fisher 1934; Yates 1934). For trigrams, TAALES calculates two versions of each of the five association measures. The first measure calculates an association score for the relationship between the first two words and the third word (e. g. *in spite* and *of*), while the second calculates an association score for the relationship between the first word and the last two words (e. g. *in* and *spite of*). For more detailed discussions of each of the bigram and trigram measures employed in this study see Evert (2009), Gries (2013), and Stefanowitsch and Gries (2003). In total, 30 indices, 10 bigram and 20 trigram indices, were included. The bigram and trigram indices were evenly split between those based on the academic subsection of COCA and the spoken subsection of COCA.

5.3 Statistical analysis

Once all texts from the three proficiency levels were analyzed by TAALES, the relevant COCA academic and spoken bigram and trigram association measure

indices were extracted. Since most of the indices were non-normally distributed, a series of Kruskall-Wallis tests were conducted to examine differences between the indices in terms of the three proficiency levels represented in the data. The indices were then checked for multi-collinearity to avoid overfitting. If any two indices were correlated at $r > 0.899$, the index with the smallest effect size was removed from subsequent analysis in order to control for multicollinearity. Any indices indicating significant differences between the three proficiency levels were then entered into a stepwise logistic regression with the purpose of determining which of these bigram and trigram association measures is most predictive of human judgments of L2 writing quality. A follow-up logistic regression with 10-fold cross validation was used to ensure generalizability of the results across the entire dataset. An alpha level of 0.0017 with a Bonferroni correction for multiple comparisons was used for all statistical tests.

6 Results

6.1 Kruskall-wallis tests

25 of the 30 bigram and trigram association measure indices included in this analysis demonstrated significant differences between proficiency levels. None of these indices were multi-collinear. The results of the Kruskall-Wallis tests, effect sizes, and descriptive statistics for these indices are presented in Table 2.

6.2 Logistic regression

The logistic regression model classified texts according to human judgments of L2 writing quality in 50.81% of cases, which is significantly higher ($df = 4$, $n = 1350$, $\chi^2 = 259.86$, $p < 0.001$, $V = 0.310$) than the baseline accuracy of 33.33%. The reported Kappa = 0.262, indicates fair agreement between actual and predicted proficiency level. The model used eight bigram and trigram association measure indices: Academic Bigram t-score, Academic Bigram ΔP, Spoken Trigram MI, Spoken Trigram ΔP, Spoken Trigram 2 Collexeme, Spoken Bigram MI^2, Academic Trigram 2 ΔP, and Spoken Trigram 2 MI^2. The simple logistic regression with 10-fold cross validation achieved a classification accuracy of 49.78%, suggesting that the predictor model was stable across the dataset. The coefficients for each of these indices are shown in Table 3. The confusion matrix for the logistic regression is shown in Table 4.

Table 2: Medians (interquartile ranges), Kruskall-Wallis χ^2 values and effect sizes for bigram and trigram association measures.

Index	A2	B1	B2	χ^2 (2, 1350)	p	η^2_p
Academic Bigram ΔP	0.034 (0.014)	0.041 (0.013)	0.046 (0.014)	206.90	< 0.001	0.15
Academic Bigram MI^2	8.53 (0.48)	8.68 (0.41)	8.81 (0.37)	166.03	< 0.001	0.12
Spoken Trigram 2 ΔP	0.117 (0.050)	0.129 (0.049)	0.145 (0.046)	120.09	< 0.001	0.09
Spoken Bigram ΔP	0.040 (0.015)	0.045 (0.014)	0.049 (0.015)	115.73	< 0.001	0.09
Academic Trigram ΔP	0.004 (0.003)	0.004 (0.003)	0.005 (0.004)	110.53	< 0.001	0.08
Academic Bigram t-score	34.89 (12.82)	38.97 (12.71)	41.43 (11.00)	109.57	< 0.001	0.08
Academic Bigram Collexeme	4824.99 (4050.82)	6353.13 (4259.41)	7163.88 (4244.85)	109.49	< 0.001	0.08
Spoken Trigram ΔP	0.004 (0.003)	0.005 (0.004)	0.006 (0.004)	101.69	< 0.001	0.08
Spoken Bigram MI	1.50 (0.31)	1.57 (0.27)	1.64 (0.25)	94.72	< 0.001	0.07
Spoken Trigram 2 MI	2.45 (0.44)	2.56 (0.42)	2.65 (0.37)	89.16	< 0.001	0.07
Spoken Trigram MI	2.46 (0.41)	2.54 (0.35)	2.64 (0.33)	86.94	< 0.001	0.06
Academic Trigram 2 ΔP	0.134 (0.054)	0.146 (0.049)	0.154 (0.049)	83.61	< 0.001	0.06
Academic Trigram Collexeme	507.14 (364.46)	556.64 (368.83)	661.34 (361.56)	60.06	< 0.001	0.04
Academic Bigram MI	1.54 (0.29)	1.59 (0.23)	1.64 (0.22)	53.99	< 0.001	0.04
Spoken Bigram MI^2	8.97 (0.56)	9.08 (0.49)	9.12 (0.42)	44.12	< 0.001	0.03
Spoken Trigram 2 Collexeme	1238.15 (1133.85)	1076.70 (794.41)	941.00 (661.61)	43.83	< 0.001	0.03
Academic Trigram 2 MI^2	8.06 (0.70)	8.12 (0.59)	8.28 (0.55)	43.19	< 0.001	0.03
Academic Trigram t-score	16.06 (5.00)	16.39 (5.00)	17.31 (4.21)	39.35	< 0.001	0.03
Academic Trigram 2 t-score	15.68 (5.34)	16.00 (4.93)	17.09 (3.96)	39.26	< 0.001	0.03
Academic Trigram MI^2	8.12 (0.71)	8.14 (0.62)	8.27 (0.51)	29.36	< 0.001	0.02
Academic Trigram 2 Collexeme	349.61 (251.51)	376.80 (224.97)	416.11 (224.18)	28.69	< 0.001	0.02
Spoken Trigram 2 MI^2	8.62 (0.65)	8.69 (0.56)	8.72 (0.49)	23.34	< 0.001	0.02
Academic Trigram 2 MI	2.72 (0.51)	2.75 (0.40)	2.81 (0.41)	17.33	< 0.001	0.01
Spoken Trigram MI^2	8.61 (0.62)	8.67 (0.48)	8.71 (0.46)	15.75	< 0.001	0.01
Spoken Trigram Collexeme	1845.34 (1823.06)	1586.59 (1372.64)	1424.74 (1202.51)	13.82	< 0.001	0.01

Table 3: Coefficients for logistic regression with 3 groups.

Index	Coefficients A2	Coefficient B1	Coefficient B2
Academic Bigram t-score	−0.11		0.10
Academic Bigram ΔP	−0.46		0.47
Spoken Trigram MI	−0.13		0.16
Spoken Trigram ΔP	−0.20		0.21
Spoken Trigram 2 Collexeme	0.23		−0.24
Spoken Bigram MI^2		0.06	
Academic Trigram 2 ΔP		0.05	
Spoken Trigram 2 MI^2		0.04	

A2 Constant: −0.10
B1 Constant: 0.10
B2 Constant: −0.09

Table 4: Confusion matrix for logistic regression with 3 groups.

	A2	B1	B2
A2	**294**	75	81
B1	177	**123**	150
B2	70	111	**269**

6.3 Post-hoc analysis

The initial analysis indicated that the logistic regression performed worst at predicting essays assigned to the B1 proficiency level (the low-intermediate level). In fact, the model only correctly predicted these texts 27.33% of the time. Thus, a post-hoc analysis was conducted that only included texts assigned to the high-beginner (A2) and high-intermediate (B2) proficiency levels. A series of Mann-Whitney U tests were conducted to determine significant differences between the two proficiency levels for the focal bigram and trigram association measure indices. Any indices showing significant differences were then entered into a simple logistic regression to determine which were most predictive of human judgments of L2 writing quality.

6.4 Mann-whitney u tests

25 of the 30 bigram and trigram association measure indices included in this analysis were significant. None of the indices were multi-collinear. The results of

Table 5: Median (interquartile ranges), Mann-Whitney U values and effect sizes for bigram and trigram association measures.

Index	A2	B2	*U* (2, 900)	p	r
Academic Bigram ΔP	0.034 (0.014)	0.046 (0.014)	156,053	< 0.001	0.47
Spoken Trigram 2 ΔP	0.117 (0.050)	0.145 (0.046)	143,708	< 0.001	0.36
Spoken Bigram ΔP	0.040 (0.015)	0.049 (0.015)	142,624	< 0.001	0.35
Academic Bigram MI^2	8.53 (0.48)	8.81 (0.37)	142,499	< 0.001	0.35
Academic Bigram t-score	34.89 (12.82)	41.43 (11.00)	141,715	< 0.001	0.35
Academic Trigram ΔP	0.004 (0.003)	0.005 (0.004)	141,470	< 0.001	0.34
Academic Bigram Collexeme	4824.99 (4050.82)	7163.88 (4244.85)	141,094	< 0.001	0.34
Spoken Trigram ΔP	0.004 (0.003)	0.006 (0.004)	139,445	< 0.001	0.33
Spoken Bigram MI	1.50 (0.31)	1.64 (0.25)	138,318	< 0.001	0.32
Spoken Trigram 2 MI	2.45 (0.44)	2.65 (0.37)	137,502	< 0.001	0.31
Spoken Trigram MI	2.46 (0.41)	2.64 (0.33)	136,469	< 0.001	0.30
Academic Trigram 2 ΔP	0.134 (0.054)	0.154 (0.049)	136,380	< 0.001	0.30
Academic Trigram Collexeme	507.14 (364.46)	661.34 (361.56)	131,076	< 0.001	0.25
Academic Bigram MI	1.54 (0.29)	1.64 (0.22)	129,027	< 0.001	0.24
Spoken Bigram MI^2	8.97 (0.56)	9.12 (0.42)	126,521	< 0.001	0.22
Academic Trigram 2 MI^2	8.06 (0.70)	8.28 (0.55)	126,086	< 0.001	0.21
Academic Trigram 2 t-score	15.68 (5.34)	17.09 (3.96)	125,225	< 0.001	0.20
Academic Trigram t-score	16.06 (5.00)	17.31 (4.21)	125,084	< 0.001	0.20
Academic Trigram 2 Collexeme	349.61 (251.51)	416.11 (224.18)	121,922	< 0.001	0.18
Academic Trigram MI^2	8.12 (0.71)	8.27 (0.51)	120,386	< 0.001	0.16
Spoken Trigram 2 MI^2	8.62 (0.65)	8.72 (0.49)	119,752	< 0.001	0.16
Spoken Trigram MI^2	8.61 (0.62)	8.71 (0.46)	116,445	< 0.001	0.13
Academic Trigram 2 MI	2.72 (0.51)	2.81 (0.41)	116,092	< 0.001	0.13
Spoken Trigram Collexeme	1845.34 (1823.06)	1424.74 (1202.51)	87,558	< 0.001	−0.12
Spoken Trigram 2 Collexeme	1238.15 (1133.85)	941.00 (661.61)	76,759	< 0.001	−0.21

the Mann-Whitney U tests, effect sizes, and descriptive statistics for these indices are presented in Table 5.

6.5 Logistic regression

The logistic regression model accurately classified texts according to human judgments of L2 writing quality in 74.89 % of cases, which is significantly higher

($df = 1$, $n = 900$, $\chi^2 = 221.02$, $p < 0.001$, $\varphi = 0.500$) than the baseline accuracy of 50 %. The reported Kappa = 0.498, indicates moderate agreement between actual and predicted proficiency level. The model used 5 association measure indices: Academic Bigram Collexeme, Academic Bigram ΔP, Spoken Trigram MI, Spoken Trigram 2 ΔP, Spoken Trigram 2 Collexeme. The simple logistic regression with 10-fold cross validation achieved a classification accuracy of 74.11 %, suggesting that the predictor model was stable across the dataset. The coefficients for each of these indices are shown in Table 6. The confusion matrix for the logistic regression is shown in Table 7.

Table 6: Coefficients for logistic regression with 2 groups.

Index	Coefficients A2	Coefficient B2
Academic Bigram Collexeme	−0.21	0.21
Academic Bigram ΔP	−0.37	0.37
Spoken Trigram MI	−0.24	0.24
Spoken Trigram 2 ΔP	−0.11	0.11
Spoken Trigram 2 Collexeme	0.30	−0.30

A2 Constant: 0.01
B2 Constant: −0.01

Table 7: Confusion matrix for logistic regression with 2 groups.

	A2	B2
A2	**338**	112
B2	114	**336**

7 Discussion

Previous association measure-based studies of n-gram use in L2 writing have shown that advanced L2 writers produce more strongly associated word pairs composed of low-frequency words, while intermediate writers rely on bigrams composed of strongly associated high-frequency words. The current study investigated whether n-gram association measures, including trigram association measures and association measures not previously employed in learner corpus studies, were predictive of human judgments of writing proficiency for beginning and intermediate L2 writers. The findings demonstrated that a logistic

regression model using eight bigram and trigram association indices could classify texts into high-beginner, low-intermediate, and high-intermediate proficiency levels with an accuracy that was significantly better than baseline. Of these eight indices, three were bigram measures and five were trigram measures. Three measures were based on the academic subsection of COCA and five were based on the spoken subsection. The post-hoc analysis, which compared indices between the high-beginner and high-intermediate groups, reported a predictive accuracy that was also significantly better than baseline using five indices of bigram and trigram association strength. Two of these indices were bigram indices based on the academic subsection of COCA, while the other three were trigram indices based on the spoken subsection of COCA. Given that these analyses only included n-gram association measures and not other measures of lexical sophistication, the findings provide support that bigrams and trigrams association strength are important factors in human judgments of L2 writing proficiency. These results have implications for our understanding of L2 phraseology and writing development, classroom instruction, and research methodologies.

The association measures found to be strongest predictors of human judgments of writing proficiency in the first logistic regression analysis were the ΔP measures. These included academic bigram ΔP (strongest), spoken trigram ΔP (third strongest), and academic trigram 2 ΔP (seventh strongest). The model indicated that texts judged as high-intermediate contained a greater number of academic bigrams and spoken trigrams that exhibited strong directional associations. Academic trigram 2 ΔP, which was only predictive for low-intermediate texts, showed that low-intermediate writers, compared to their high-beginner counterparts, employed more academic trigrams with strong directional associations between the first word and the following two. The second strongest overall predictor was spoken trigram 2 collexeme, indicating that higher scoring texts contained fewer spoken trigrams with strong associations (according to the collexeme measures in TAALES) between the first word and the last two words in the trigram. Spoken trigram MI was found to be the fourth strongest predictor of writing proficiency. This indicated that students who wrote higher scoring essays employed more strongly associated spoken trigrams that were composed of lower frequency words. Comparing the trends for spoken trigram 2 collexeme and spoken trigram MI, the findings suggest that while the higher scoring essays contained less strongly associated spoken trigrams overall, their spoken trigrams composed of low-frequency words showed stronger associations. The fifth strongest predictor of human judgments of writing proficiency was academic bigram t-score. The model showed that writers of high-intermediate texts employed more strongly associated academic bigrams composed of

high-frequency words. Last, the logistic regression indicated that spoken bigram MI^2 and spoken trigram 2 MI^2 were the sixth and eighth strongest predictors of L2 writing quality. These indices, which were only predictive for low-intermediate texts, indicated that writers of low-intermediate texts employed more strongly associated spoken bigrams and trigrams compared to their high-beginner counterparts. Overall, this analysis showed that the more proficient writers in this study employed more strongly associated academic bigrams as well as more strongly associated spoken trigrams.

Three indices found to be predictive of writing quality in the original analysis were also found to be predictive in the post-hoc analysis, which only included high-beginner and high-intermediate texts. These included academic bigram ΔP, spoken trigram 2 collexeme, and spoken trigram MI. Similar to the original analysis, academic bigram ΔP was found to be the strongest predictor in the post-hoc analysis, indicating that students who employed more academic bigrams that are stronger in their directional association were more likely to have their texts classified as high-intermediate. Spoken trigram 2 collexeme was the second strongest predictor of human judgments of writing proficiency, indicating that high-intermediate texts contained fewer strongly associated trigrams (as measured by the Fisher-Yates test) common to native speech. Spoken trigram MI was the third strongest predictor in the post-hoc analysis. The model showed that, in contrast to the results for spoken trigram 2 collexeme, writers of the high-intermediate texts employed more strongly associated spoken trigrams that were composed of low-frequency words. The fourth strongest predictor in the post-hoc analysis was academic bigram collexeme. The logistic regression indicated that students receiving higher scores on their essays employed more strongly associated bigrams (calculated by the Fisher-Yates test) common to academic writing. The fifth strongest predictor was spoken trigram 2 ΔP, indicating that the writers of the high-intermediate texts employed more spoken trigrams that showed a strong directional association between the first word and last two words in the trigram. Similar to the first analysis, this analysis found that intermediate writers, compared to their beginner counterparts, produced more strongly associated academic bigrams and spoken trigrams in their placement essays.

While not included in either the original or post-hoc regression models, several other indices showed significant differences between texts at the three proficiency levels. These differences generally support the findings from the models. First, all six ΔP indices showed significant differences between the three groups, indicating that higher scoring essays contained bigrams and trigrams that had stronger directional associations. These indices also had the highest average effect size for all association measures in the analyses

($\eta_p^2 = 0.091$ and $r = 0.359$ respectively). Five MI and six MI^2 indices also showed significant differences between the proficiency levels, indicating that high-scoring essays contained more strongly associated spoken and academic bigrams and trigrams composed of low-frequency words. The MI indices had the second highest average effect size in both analyses ($\eta_p^2 = 0.051$ and $r = 0.258$ respectively) with MI^2 having the third highest ($\eta_p^2 = 0.040$ and $r = 0.205$ respectively). The results for the collexeme measures showed that while academic bigram, academic trigram, and academic trigram 2 collexeme scores were significantly higher for high-intermediate texts, scores for spoken trigram collexeme and spoken trigram 2 collexeme scores were significantly lower. Lastly, only three t-score measures showed significant differences between L2 writers at different proficiency levels. These three indices, which were the weakest in terms of average effect size, showed that high-intermediate writers employed more academic bigrams and trigrams composed of high-frequency words.

Taken together, these results show that L2 writers judged to be more proficient have a greater productive knowledge of bigrams and trigrams composed of strongly associated words. They specifically point to three key differences between more and less proficient L2 writers. First, the inclusion of two academic bigram indices in both logistic regression analyses indicates that more proficient writers employ a greater number bigrams characteristic of academic writing. This finding supports results from Chen and Baker (2014), who found that more advanced L2 writers produced more 4-grams characteristic of academic prose. Second, the results from the academic bigram ΔP index, which was the strongest predictor in both analyses, suggests that a greater productive knowledge of academic bigrams includes a greater understanding of the preferred sequences for co-occurring words (i. e. which comes first in combination). Lastly, the results for the MI measures indicate that more proficient L2 writers employ a greater number of bigrams and trigrams composed of strongly associated words that occur less frequently in native speech. This finding, along with those from previous studies of lexical sophistication and L2 writing (Crossley and McNamara 2012; Crossley et al. 2014) suggests that, as learners develop their writing proficiency, they not only increase their knowledge of low-frequency words, but also develop a better understanding of the associations these words share with one another. More proficient writers are then better able to employ this knowledge in their writing, combining low-frequency words together into more cohesive multi-word sequences. This supports results from Granger and Bestgen (2014), who found that advanced L2 writers employed more strongly associated, low-frequency bigrams than intermediate writers. Thus, between the current study and Bestgen and Granger (2014), a clear trend can be noticed for bigram use by L2 writers such that as proficiency increases, L2 writers'

knowledge of these low-frequency bigrams also appears to increase. Findings from this study also points to similar trends for trigrams, although more research is needed with advanced writers.

Despite the strong results of this study, it is important to note that the original regression model performed poorly in predicting texts judged as low-intermediate. The reason for this may be that development of strongly associated bigrams and trigrams may take longer to develop than other types of phraseological knowledge. Such knowledge may also take longer to develop than other types of linguistic knowledge. Indeed, the longitudinal study by Bestgen and Granger (2014) found that, even after a semester of intensive study, ESL writers made little improvement in their use of strongly associated bigrams. Thus, while the association strength of bigrams and trigrams may be a good measure of L2 writing proficiency across proficiency levels one step apart (i. e. high-beginner compared to high-intermediate), it may be not strong enough to differentiate between adjacent proficiency levels.

In terms of pedagogical implications, these results support increased attention on multi-word sequences in foreign language classrooms. They also indicate that attention should be given to certain kinds of n-grams as well as information regarding the nature of n-grams. According to Gries and Ellis (2015), the learning of multi-word sequences in English requires a massive amount of exposure to the target language, something EFL learners often have difficulty attaining. If L2 learners are going to be better prepared to use multi-word sequences in their writing, increased exposure through direct instruction focusing on collocations, lexical bundles, and other multi-word sequences is needed. In terms of which sequences to include in such instruction, the results of the current study highlight indicate that attention be given to sequences composed at least in part of low-frequency words. Because these items are more predictive of human ratings, knowledge of these items may help L2 writers improve their writing skills. In addition, the results regarding ΔP indicate that it may not be enough for learners to know that two or three words are associated. They may also need to be aware of how associated words are sequenced. The Academic Formulas List (AFL; Simpson-Vlach and Ellis 2010) and the Academic Collocations List (ACL; Ackerman and Chen 2013) are two examples of pedagogically motivated attempts that employ MI scores in the selection of multi-word sequences. Future enhancement of these lists and other approaches to identifying pedagogically relevant multi-word sequences could also make use of ΔP to tease apart the directional association of n-grams identified through frequency counts, MI or t-score. For example, a list of target n-grams could be extracted from a reference corpus and then refined through the use of ΔP. This refinement could include identifying the preferred order of strongly associated words

combinations or reducing the number of target n-grams to those showing the strongest directional associations. Another possibility would be determining which word is the stronger cue for the other and then presenting the n-grams to the learners in that sequence (i. e. first presenting the word that is the strongest cue for the second).

Lastly, the results of this study also have implications for the selection of n-gram association measures in future learner corpus studies. First, the results regarding ΔP support claims by Gries (2013) and Ellis and Gries (2015) for the inclusion of this measure in phraseological analyses. ΔP can reveal more about the extent to which learners understand the association between words (i. e. not just strength, but order as well). In addition, because ΔP measures contingency between cues and outcomes (in the case of n-grams, words) and not simple association, this measure can more accurately reflect a language learner's experience with the target language. Second, the results highlight the utility of n-gram indices based on reference corpora representing different registers. The results indicate that by including indices based on spoken and academic registers, researchers can investigate how learners shift their use of n-grams characteristic of different registers as their proficiency increases.

8 Conclusion

The results of the current study indicate that indices related to bigram and trigram use can be used to predict human judgments of L2 writing quality for beginner and intermediate writers. L2 writers rated as intermediate by human raters were found to employ more strongly associated academic bigrams and spoken trigrams compared to writers classified as beginners. In addition, through the use of multiple types of association measures, the current study has shown that computational indices accounting for the directionality of word association may provide more valid measures of L2 n-gram knowledge than symmetric indices. These results have important implications for L2 writing pedagogy as well as future studies of L2 phraseological knowledge.

Future studies examining the connection between n-gram association measures and L2 writing proficiency should include larger and more diverse learner corpora. These corpora could include learners from more diverse L1 backgrounds, from ESL as well as EFL contexts, and texts from a wider range of writing tasks. They could also include texts from advanced learners, thus allowing a more complete view of n-gram use as it related to L2 writing proficiency. Future research could also examine the development of productive

n-gram knowledge through the use of small-scale longitudinal learner corpus data. This research could provide information regarding the dynamic process through which L2 productive phraseological knowledge develops for individual learners.

References

Ackerman, K. & Y.H. Chen. 2013. Developing the Academic Collocations List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12. 235–247.

Ädel, A. & B. Erman. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31. 81–92.

Bestgen, Y. & S. Granger. 2014. Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing* 26. 28–41.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Bialystok, E. 1978. A theoretical model of second language learning model. *Language Learning* 28. 69–83.

Chen, Y.H. & P. Baker (2014). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2, C1. *Applied Linguistics Advance Access published December 5, 2014*. 10.1093/applin/amu065

Crossley, S.A. 2013. Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching* 46(2). 256–271.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. 2011. Predicting lexical proficiency in language learners using computational indices. *Language Testing*, 28, 561–580.

Crossley, S.A., Z. Cai & D.S. McNamara (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. M. McCarthy & G. M. Youngblood (eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 214–219. Menlo Park, CA: The AAAI Press.

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

Crossley, S.A., K. Kyle, L.K. Allen, L. Guo & D.S. McNamara. 2014. Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation. *The Journal of Writing Assessment* 7. 1.

Crossley, S.A., K. Kyle & D.S. McNamara. 2016. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *The Journal of Second Language Writing* 32. 1–16.

Crossley, S.A. & D.S. McNamara. 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. & D.S. McNamara. 2013. Applications of text analysis tools for spoken response grading. *Language Learning and Technology* 17(2). 171–192.

Crossley, S.A. & D.S. McNamara. 2014. Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing* 26(4). 66–79.

Davies, M. (2009). The Corpus of Contemporary American English: 450 million words, 1990-present. Avaliable online at: http://corpus/byu.edu/coca

DeCock, S., S. Granger, G. Leech & T. McEnery. 1998. An automated approach to the phrasicon of EFL learners. In S. Granger (ed.), *Learner English on Computer*, 67–79. London: Longman.

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

Ellis, N.C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.

Ellis, N. C., & Simpson-Vlach, R. 2009. Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory* 5(1). 61–78.

Ellis, N.C. 2012. Formulaic language and second language acqusition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics* 32. 17–44.

Erman, B., & Warren, B. 2000. The idiom principle and the open choice principle. *Text*, 20. 29–62.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. (Dissertation). University of Stuttgart: Stuttgart.

Evert, S. 2009. Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, 1211–1248. Berlin: Mouton de Gruyter.

Fisher, R.A. 1934. *Statistical Methods for Research Workers*. 2nd edn. Edinburgh: Oliver and Boyd.

Friginal, E., M. Li & S. Weigle. 2014. Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing* 23. 1–16.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*, 145–160. Oxford: Oxford University Press.

Granger, S. & Y. Bestgen. 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.

Granger, S. & M. Paquot. 2008. Disentangling the phraseological web. In S. Granger & F. H. Meunier (eds.), *Phraseology: An Interdisciplinary Perspective*, 27–49. Amsterdam: John Benjamins.

Grant, L. & A. Ginther. 2000. Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing* 9(2). 123–145.

Gries, S.T., Hampe, B., & Schönfeld, D. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.

Gries, S.T. 2013. 50-something years of work on collocations: What is or should be next …. *International Journal of Corpus Linguistics* 18(1). 137–165.

Gries, S.T. & N.C. Ellis. 2015. Statistical measures for Usage-Based Linguistics. *Language Learning* 65(Supplement 1). 228–255.

Groom, N. 2009. Effects of second language immersion on second language collocational development. In A. Barfield & H. Gyllstad (eds.), *Researching Collocations in Another Language: Multiple Interpretations*, 21–33. Bastingstoke: Palgrave Macmillan.

Guo, L., S.A. Crossley & D.S. McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparative study. *Assessing Writing* 18. 218–238.

Hyland, K. 2008. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18(1). 41–62.

Ishikawa, S. 2009. Phraseology overused and underused by Japanese learners of English: A contrastive Interlanguage Analysis. *Phraseology, Corpus Linguistics and Lexicography*: Papers from Phraseology 2009 in Japan, 87–100. Nishinomiya: Kwansei Gakuin University Press.

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

Jin, W. 2001. A quantitative study of cohesion in Chinese graduate students' writing: Variations across genres and proficiency levels. ERIC database (ED452726). (accessed).

Kyle, K. & S.A. Crossley. 2015. Automatically assessing lexical sophsitication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4). 757–786.

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

Laufer, B. & P. Nation. 1995. Vocabulary size and use: Lexical Richness in L2 Written Production. *Applied Linguistics* 16(3). 307–322.

Meara, P. 2005. Designing vocabulary tests for English, Spanish and other languages. In C. Butler, S. Christopher, M.A. Gómez González, & S. M. Doval-Suárez (Eds.), *The Dynamics of Language Use* (pp. 271–285). Amsterdam: John Benjamins.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24(2). 223–242.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Nunan, D. 1989. *Designing tasks for the classroom*. Cambridge: Cambridge University Press.

Ohlrogge, A. 2009. Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, E. A. Moravcsik, H. Ouali & K. M. Wheatley (eds.), *Formulaic Language (Volume 2): Acquisition, Loss, Psychological Reality, and Functional Explanations*, 375–385. Amsterdam: John Benjamins.

Reid, J. 1992. A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing* 1. 79–107.

Rhee, S.C. & C.K. Jung. 2014. Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *The Journal of the Korea Contents Association* 14(11). 1019–1029.

Read, J. 2000. Assessing Vocabulary. Cambridge: Cambridge University Press.

Read, J. 2004. Plumbing the depths: How should the construct of vocabulary knowledge be defined. Vocabulary in a Second Language: Selection, Acquisition, and Testing, 10(1). 209–227.

Römer, U. 2009. The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Applied Linguistics* 7. 140–162.

Simpson-Vlach, R. & N.C. Ellis. 2010. An academic formulas list (AFL). *Applied Linguistics* 31. 487–512.

Siyanova-Chantura, A. & R. Martinez. 2015. The idiom principle revisited. *Applied Linguistics* 36(5). 549–569.

Stefanowitsch, A. & S.T. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.

Vidakovic, I. & F. Barker. 2010. Use of words and multi-word units in Skills for Life Writing examinations. *Cambridge ESOL: Research Notes* 41. 7–14.
White, R. 1981. Approaches to writing. *Guidelines* 6. 1–11.
Yates, F. 1934. Contingency tables involving small numbers and the χ2 test. *Journal of the Royal Statistical Society, Supplement* 1. 217–235.