# Predicting L2 Writing Proficiency with Computational Indices Based on N-grams

Byung-Doh Oh
(Seoul National University)

**Oh, Byung-Doh. (2017). Predicting L2 writing proficiency with computational indices based on n-grams. *Foreign Language Education Research, 21,* 1-20.**

Linguistic features that are indicative of higher writing proficiency levels can inform many aspects of language assessment such as scoring rubrics, test items, and automated essay scoring (AES). The recent advancement of computer algorithms that automatically calculate indices based on various linguistic features has made it possible to examine the relationship between linguistic features and writing proficiency on a larger scale. While the ability to use appropriate n-grams – recurring sequences of contiguous words – has been identified as a characteristic differentiating between proficiency levels in the literature, few studies have examined this relationship using computational indices. To this end, this study utilized the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) to calculate eight indices based on n-grams from a stratified corpus consisting of 360 argumentative essays written by Korean college-level learners. First, the indices from the training set of 240 essays were used to design a multinomial logistic regression model in order to identify indices that are significant predictors of writing proficiency levels. Subsequently, the regression model was applied to a test set of 120 essays to examine whether the model could be used to predict the proficiency levels of unseen essays. The results revealed that the *mean bigram T*, *mean bigram Delta P*, *mean bigram-to-unigram Delta P*, and *proportion of 30,000 most frequent trigrams* indices were significant predictors of proficiency levels. Furthermore, the regression model based on eight indices correctly classified 52.5% of essays in the test set, demonstrating above-chance level accuracy.

**Key Words**: L2 writing proficiency, n-grams, phraseology, computational linguistics, language assessment

## Ⅰ. Introduction

A prominent issue in language assessment is identifying linguistic features that are predictive of higher proficiency levels. Pinpointing exactly what comprises high-quality writing can influence many aspects of language testing and assessment, including the structure of scoring rubrics (Hawkins & Filipović, 2012), development of test items (Barker, Salamoura, & Saville, 2015), and selection of features for automated essay scoring (AES; Crossley, Kyle, Allen, Guo, & McNamara, 2014). To this end, many studies have analyzed the relationship between linguistic features and proficiency level using stratified learner corpora. While this has been a research topic since the 1970s (see Wolfe-Quintero, Inagaki, & Kim, 1998, for overview), the recent development of various computer algorithms has made it possible to examine this relationship on a larger scale. These computer algorithms employ natural language processing (NLP) tools such as tokenizers, part-of-speech taggers, and parsers (Jurafsky & Martin, 2008) as well as frequency data from other corpora to automatically process the input text and generate indices based on different linguistic features. Some widely known programs open to the public include Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014), L2 Syntactic Complexity Analyzer (Lu, 2010), and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015).

An important component of writing ability is using words appropriately together in context, which is the object of study in phraseology (Ebeling & Hasselgård, 2015). While there is no consensus on how to operationalize such co-occurrence of words, one of the main strands of phraseology is the study of n-grams. More commonly referred to as *lexical bundles*[1] in the context of learner corpus research, they are defined as the most frequently recurring sequences of contiguous words, regardless of their idiomaticity and structural status (Biber, Johansson, Leech, Conrad, & Finegan, 1999). Because of this minimal constraint, although lexical bundles can be easily identified within a text, they pose challenges in terms of their linguistic and qualitative interpretation (Ebeling & Hasselgård, 2015). Nonetheless, previous studies on lexical bundles have revealed significant findings regarding language use across different registers, and consequently lexical bundles have been recognized as important building blocks of discourse (Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004; Hyland, 2008). This insight is recently being applied to the field of learner corpus research to examine the relationship between n-gram use and writing proficiency levels from different perspectives. More specifically, previous studies have analyzed frequent n-grams to identify stylistic differences (Chen & Baker, 2016; Staples, Egbert, Biber, & McClair, 2013) and compared n-grams in learner writing to n-grams in a representative native speaker reference corpus in terms of frequency and overlap (Crossley, Cai, & McNamara, 2012), as well as collocational

1 Although n-grams and lexical bundles are generally regarded as synonymous in the literature, many studies on lexical bundles adopt stricter criteria for their operationalization, i.e., a minimum cut-off frequency and a minimum number of different texts they need to occur in. Furthermore, while n-grams often refer to two-word sequences (bigrams) or three-word sequences (trigrams) that easily lend themselves to computational processing due to their high frequency, lexical bundles commonly denote four-word sequences that may be less frequent but demonstrate a higher degree of syntactic/pragmatic completeness. In this article, the two terms are used in accordance with the specific study cited.

strength (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014). This study aims to extend this line of research by incorporating more computational indices related to n-gram use and examining their potential to predict proficiency levels of unseen essays.

## Ⅱ. Literature Review

In the literature, various analyses have been conducted with indices calculated on learner writing and holistic ratings assigned by human raters. Many previous studies have tried to identify indices that are significant predictors of writing proficiency, by means of statistical analyses such as multiple regression and discriminant function analysis. For example, McNamara, Crossley, and McCarthy (2010) examined the relationship between 26 Coh-Metrix indices and the holistic scores of 120 argumentative essays written by college freshmen. While they observed no correlation between measures of cohesion and writing quality, they revealed that indices representing syntactic complexity, lexical diversity, and word frequency were significantly correlated to and could predict writing quality. The relationship between textual cohesion and writing quality was found to be actually negative by Crossley and McNamara (2012), who conducted a similar study with 514 essays written as responses to the Hong Kong Advanced Level Examination (HKALE). They revealed that while essays judged as more proficient contained less cohesive devices, they demonstrated a higher level of linguistic sophistication in terms of lexical diversity, word frequency, word meaningfulness, and word familiarity. Similarly, Kim (2014) utilized the Lexical Complexity Analyzer (Lu, 2012) and the L2 Syntactic Complexity Analyzer (Lu, 2010) to examine the argumentative writing of college-level Korean EFL learners across different proficiency levels. She also discovered that some indices representing text length, lexical complexity, and syntactic complexity were predictive of proficiency levels.

More relevant to the present research are studies that compared lexical bundles in texts of different proficiency levels to discover noteworthy differences in their use. For instance, Staples et al. (2013) analyzed four-word lexical bundles in 960 responses to the TOEFL iBT writing section categorized into three groups according to their scores. Examination of the bundles across the three levels revealed that the highest scoring responses contained less repetitive lexical bundles, including those influenced directly by the task prompt. Similarly, Chen and Baker (2016) graded essays in the Longman Learner Corpus (LLC) according to the Common European Framework of Reference for Languages (CEFR) and categorized four-word lexical bundles in essays that were rated B1, B2, and C1. They identified a stylistic difference in the use of lexical bundles, i.e., less proficient writing sharing more features with conversation and more proficient writing demonstrating a more impersonal and academic tone.

Other studies have tried to capture this qualitative difference by comparing n-grams in learner writing to those found in a representative native corpus like the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA). Crossley et al. (2012), for instance, developed a set of algorithms to quantify the accuracy, frequency, and proportion of n-grams in learner writing by utilizing n-gram frequency data from the BNC. A multiple regression analysis with these indices calculated from 313 college-level essays and their holistic ratings revealed that the holistic ratings were negatively correlated to n-gram proportion indices and frequency indices. In other words, essays that were rated as higher quality contained less n-grams that were found in the BNC.

Another method of operationalizing n-gram use in learner writing is by means of measuring their association strength, i.e., how much more a sequence of words is likely to co-occur than by chance. By assigning each n-gram with association measures calculated from a representative reference corpus, the association strength of n-grams in learner writing can be compared. For example, Durrant and Schmitt (2009) calculated the *t*-score and mutual information (MI) based on the BNC for adjective-noun and noun-noun bigrams found in native writing and non-native writing. They found that while non-native writers overused frequently occurring bigrams identified by high *t*-scores, they underused strong collocations that are characterized by high MI. Similar findings were observed in a later study by Granger and Bestgen (2014), who extended Durrant and Schmitt's (2009) methodology to other types of bigrams (i.e., adverb-adjective and 'all' regardless of part of speech). Comparison of the bigrams in 223 CEFR-graded essays from the International Corpus of Learner English (ICLE) confirmed a similar relationship between writing proficiency and *t*-score/MI. Additionally, Bestgen and Granger (2014) incorporated another index into their analysis of bigrams in the Michigan State University corpus – namely, the proportion of bigrams absent in the reference corpus. Their cross-sectional study revealed that the rated quality of text was positively correlated to mean MI but negatively correlated to the proportion of bigrams absent in the reference corpus.

Despite such salient difference in the characteristics of n-grams used across different proficiency levels, few studies have attempted to examine this relationship using a number of computational indices that comprehensively reflect different aspects of n-gram use (e.g., frequency, proportion, collocational strength). Furthermore, to our knowledge, no study has been conducted to examine whether the proficiency level of unseen essays can be predicted based on these indices. In light of the discussion so far, the aim of the present study is to address the two aforementioned issues surrounding the relationship between computational indices and writing proficiency – identifying significant predictors and making predictions based on them.

This study was initiated to answer the following two questions:

(1) Which indices based on n-grams are significant predictors of writing proficiency of Korean EFL learners?

(2) To what extent can indices based on n-grams predict the writing proficiency of Korean EFL learners?

# Ⅲ. Methods

## 1. Corpus

### 1.1. Learner Corpus

The corpus used in this study is a sub-corpus of 360 argumentative essays from the Yonsei English Learner Corpus (YELC; Rhee & Jung, 2014). YELC consists of a total of 6,572 essays (3,286 narrative and 3,286 argumentative) written by Korean college-level learners of English (and those with similar qualifications) that were admitted to Yonsei University in 2011. As a part of the computer-based Yonsei English Placement Test (YEPT), these college-level learners were asked to write a narrative essay about 100 words long on a familiar topic and an argumentative essay about 300 words long on an academic topic[2]. The learners were given 60 minutes to complete a word rearranging task as well as the narrative and argumentative essays. All essays were graded by trained native speakers, and a holistic proficiency level was assigned to each learner based on the grades. There are a total of nine proficiency levels, which resulted from the calibration of the CEFR to the nine-band grading scale of the Korean College Scholastic Aptitude Test (CSAT).

It should be noted that this proficiency level is not solely based on the argumentative essay due to the fact that one grade was assigned to each learner based on the results of the entire YEPT. Furthermore, because the grading scale of CSAT is norm-referenced in nature, the validity of aligning the CEFR to these nine grades is also questionable. Despite this shortcoming, the YELC was selected as the learner corpus of this study due

2 Although the creators of the YELC did not disclose the writing prompts used on the YEPT, according to Choe and Song (2013), there are a total of six writing topics for the argumentative essays. They are; (1) physical punishment in schools, (2) using animals in medical experiments, (3) smoking in public buildings, (4) using cellular phones while driving, (5) compulsory military service, and (6) using real names on the Internet.

to its homogeneity in terms of the learners' language background (i.e., Korean) and the nature of the writing task. The constrained testing situation of the YEPT resulted in essays that are generally comparable and differentiated only by writing proficiency, which is the main focus of the present study.

For this study, essays under the nine proficiency levels were classified into three different proficiency groups (i.e., A1, A1+, and A2 into basic level, B1, B1+, and B2 into intermediate level, and B2+, C1, and C2 into advanced level). The advanced level group had the smallest number of 120 essays, due to the fact that students that had already acquired a high proficiency level of English were exempt from the YEPT (Rhee & Jung, 2014). In order to retain as much data as possible yet avoid overrepresentation of a certain proficiency group, the same number of 120 essays were randomly sampled from both the basic level group and the intermediate level group. In the process, only essays longer than 100 words were selected, as essays shorter than 100 words are not suitable for the calculation of automated indices (Crossley & McNamara, 2013) and do not contain enough n-grams for analysis. The number of essays and tokens from each proficiency group is summarized in Table 1.

**TABLE 1**

Classification used in the current study

| | Basic | Intermediate | Advanced | Total |
|---|---|---|---|---|
| Essays | 120 | 120 | 120 | 360 |
| Tokens | 25,115 | 30,407 | 36,265 | 91,787 |
| Mean number of tokens | 209.29 | 253.39 | 302.21 | 254.96 |

### 1.2. Reference Corpus

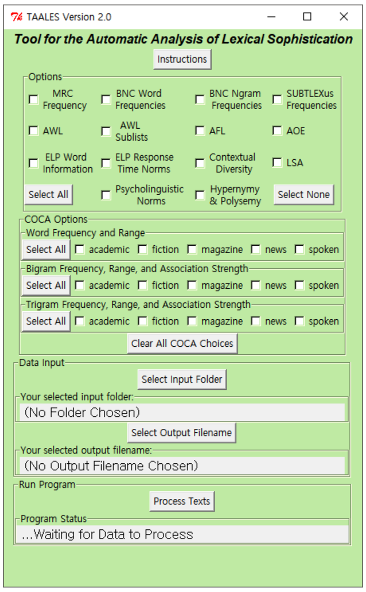
Some indices based on n-grams analyzed in the current study rely on frequency data from a reference corpus. TAALES, the text analysis tool primarily used in the current study, offers a range of five reference corpora to choose from, which are the five sub-corpora of COCA (i.e., academic, fiction, magazine, news, and spoken). Although none of the sub-corpora of COCA directly contains argumentative writing explicitly expressing one's opinion, the academic sub-corpus was chosen as the reference corpus due to its similarity in style (e.g., degree of formalness) with the argumentative essays of YELC. Furthermore, the representative nature of COCA was thought to shed light on n-grams native speakers of English commonly use in writing.

## 2. Tools

To automatically calculate various indices based on n-grams for each essay in the learner corpus, TAALES Version 2.0 (Figure 1) was utilized. Among the indices TAALES offers, 43 indices are linked to n-gram use (see Appendix 1). There are two types of indices based on n-grams – mean indices and proportion indices. For mean indices, TAALES automatically processes the input text to identify n-grams that occur in the reference corpus. Then it assigns each n-gram its relevant score (e.g., bigram *t*-score, bigram MI) calculated based on frequency data from the reference corpus. Finally, the sum of scores is divided by the number of n-grams that was assigned a score, which results in a mean index (e.g., mean bigram *t*-score, mean bigram MI) of each input text. On the other hand, proportion indices are calculated by dividing the number of n-grams that overlap with those that occur in the reference corpus by the total number of n-grams within the text. Therefore, while n-grams that do not occur in the reference corpus do not affect mean indices, they lower proportion indices as they are included in the total number of n-grams[3].

**FIGURE 1**

User interface of TAALES Version 2.0




3 For example, the two strings *the nicest person I know is* and *the nicest person I know is Gildong Hong* would yield the same mean bigram indices but different bigram proportion indices, if *is Gildong* and *Gildong Hong* do not occur in the reference corpus.

## 3. Statistical Analyses

Following previous studies in the literature (Crossley & McNamara, 2012; Crossley et al., 2014; Jung, Crossley, & McNamara, 2015; Kim, 2014; McNamara et al., 2010), the essays were first randomly split into a training set (67%, 240 essays) and a test set (33%, 120 essays). The training set was used to design a multinomial logistic regression model[4] with indices based on n-grams as independent variables and proficiency group as the dependent variable. Subsequently, the regression model was used to predict the proficiency groups of the 120 essays in the test set. The main reason for separating a portion of the essays into a test set was to examine the generalizability of the regression model – i.e., how well it can predict the proficiency groups of similar essays that were not included in the training data. All statistical analyses were conducted using the Statistical Package for the Social Sciences (SPSS) Version 24.0.

After the 43 indices for each essay in the training set have been calculated, a series of statistical analyses were conducted to finalize the variables for the regression model. First, a one-way analysis of variance (ANOVA) was conducted to compare the means of each group. As mentioned earlier, because most indices did not satisfy the underlying assumptions of normality and homoscedasticity, Welch's correction for heteroscedasticity was applied[5]. Two indices that did not significantly differ across proficiency groups were excluded from subsequent analyses (see Appendix 2 for the descriptive statistics and ANOVA results for all indices). Next, in order to prevent multicollinearity, any two indices that showed a Pearson's $r$ higher than or equal to 0.7 (Crossley & McNamara, 2012; Crossley et al., 2014; McNamara et al., 2010) were initially flagged. Then, indices with a smaller $F$ value from the initial ANOVA were removed until none of the indices demonstrated a strong correlation ($r \geq 0.7$) with each other. This resulted in a total of eight independent variables to be included in the regression model (Table 2).

4 While previous studies used discriminant function analysis and multiple regression for the same purpose, preliminary screening of the data revealed that many of the indices based on n-grams violated the assumptions of normality and homoscedasticity. Therefore, multinomial logistic regression, which is robust against the violation of these assumptions, was selected instead.

5 Studies like Staples et al. (2013) opted for the non-parametric Kruskall-Wallis ANOVA because the two underlying assumptions were not met. However, because the Kruskall-Wallis ANOVA also assumes homoscedasticity, Welch's ANOVA is more appropriate for non-normal, heteroscedastic data (McDonald, 2014).

**TABLE 2**

Independent variables of the regression model

| Variable | Formula |
|---|---|
| Mean bigram *T* (bi_T) | $\frac{p(w_1,w_2)-p(w_1)p(w_2)}{\sqrt{\frac{p(w_1,w_2)}{N}}}$ ($N$ = total tokens) |
| Mean bigram-to-unigram *T* (tri_2_T) | |
| Mean bigram MI (bi_MI) | $\log\frac{p(w_1,w_2)}{p(w_1)p(w_2)}$ |
| Mean bigram-to-unigram MI$^2$ (tri_2_MI2) | $\log\frac{p(w_1,w_2)^2}{p(w_1)p(w_2)}$ |
| Mean bigram Delta P (bi_DP) | $p(w_2\|w_1)-p(w_2\|\neg w_1)$ |
| Mean unigram-to-bigram Delta P (tri_DP) | |
| Mean bigram-to-unigram Delta P (tri_2_DP) | |
| Proportion of 30,000 most frequent trigrams (tri_prop_30k) | $\frac{\text{Number of n-grams in reference corpus}}{\text{Total number of n-grams}}$ |

*T*-score, MI, MI$^2$, and Delta P are measures that are used to calculate the association strength of bigrams. *T*-score is calculated by applying the statistical *t*-test to bigrams, to examine whether the probability of a bigram occurring is significantly higher than the product of the probabilities of its individual words occurring. MI refers to the amount of information gained about the occurrence of a word at position *i* once aware of the word at position *i+1*, and vice-versa (Manning & Schütze, 1999). The main difference between the two statistical measures is that while "rankings based on *t*-scores tend to highlight very frequent collocations […], MI tends to give prominence to word pairs which may be less common, but whose component words are not often found apart" (Durrant & Schmitt, 2009, p. 167). MI$^2$ is a variant of MI developed to mitigate such overestimation of low-frequency pairs (Evert, 2005). The Delta P score reflects the probability of an outcome (i.e., a particular word) based on a cue (i.e., another word) and is calculated by subtracting from the probability of an outcome given a cue the probability of an outcome without the cue (K. Kyle, personal communication, December 15, 2016). While these measures of association are not traditionally calculated for trigrams, TAALES utilizes two different methods to calculate them; i.e., by treating the first two words of a trigram as a single unit (bigram-to-unigram) and by treating the last two words of a trigram as a single unit (unigram-to-bigram).

## IV. Results and Discussion

### 1. RQ 1: Significant Predictors of Writing Proficiency

A multinomial logistic regression analysis using eight indices as independent variables yielded a significant statistical model, $\chi^2(16) = 111.972$, $p < 0.01$. The likelihood ratio tests of the regression model identified four variables as significant predictors of proficiency level: *mean bigram T*, *mean bigram Delta P*, *mean bigram-to-unigram Delta P*, and *proportion of 30,000 most frequent trigrams*. Furthermore, pairwise comparisons of the adjacent proficiency groups revealed that the regression coefficients for *mean bigram-to-unigram Delta P* and *proportion of 30,000 most frequent trigrams* were significant between the basic and intermediate groups, while the regression coefficients for *mean bigram T*, *mean bigram Delta P*, and *proportion of 30,000 most frequent trigrams* were significant between the intermediate and advanced groups. The results are summarized in Tables 3 and 4.

**TABLE 3**

Likelihood ratio tests of the regression model

| Variable | -2log-likelihood of reduced model | $\chi^2$ | Significance |
|---|---|---|---|
| Mean bigram *T* (bi_T) | 422.603 | 7.543 | 0.023* |
| Mean bigram-to-unigram *T* (tri_2_T) | 415.141 | 0.081 | 0.960 |
| Mean bigram MI (bi_MI) | 417.551 | 2.512 | 0.285 |
| Mean bigram-to-unigram MI$^2$ (tri_2_MI2) | 415.919 | 0.860 | 0.651 |
| Mean bigram Delta P (bi_DP) | 422.242 | 7.183 | 0.028* |
| Mean unigram-to-bigram Delta P (tri_DP) | 415.645 | 0.586 | 0.746 |
| Mean bigram-to-unigram Delta P (tri_2_DP) | 422.740 | 7.681 | 0.021* |
| Proportion of 30,000 most frequent trigrams (tri_prop_30k) | 440.752 | 25.693 | 0.000** |

Significance level: * $p < 0.05$, ** $p < 0.01$

**TABLE 4**

Regression coefficients between adjacent proficiency groups

| Comparison group | Variable | B | Wald | Significance |
|---|---|---|---|---|
| Basic | Mean bigram *T* (bi_T) | 0.002 | 0.010 | 0.920 |
| | Mean bigram-to-unigram *T* (tri_2_T) | -0.018 | 0.079 | 0.778 |
| | Mean bigram MI (bi_MI) | -0.798 | 0.550 | 0.458 |
| | Mean bigram-to-unigram MI$^2$ (tri_2_MI2) | -0.273 | 0.260 | 0.610 |
| | Mean bigram Delta P (bi_DP) | -5.035 | 0.034 | 0.854 |
| | Mean unigram-to-bigram Delta P (tri_DP) | -0.985 | 0.000 | 0.990 |
| | Mean bigram-to-unigram Delta P (tri_2_DP) | -16.189 | 5.965 | 0.015* |
| | Proportion of 30,000 most frequent trigrams (tri_prop_30k) | -16.083 | 6.617 | 0.010* |
| Advanced | Mean bigram *T* (bi_T) | -0.057 | 6.059 | 0.014* |
| | Mean bigram-to-unigram *T* (tri_2_T) | -0.009 | 0.013 | 0.908 |
| | Mean bigram MI (bi_MI) | 1.233 | 1.166 | 0.280 |
| | Mean bigram-to-unigram MI$^2$ (tri_2_MI2) | -0.594 | 0.827 | 0.363 |
| | Mean bigram Delta P (bi_DP) | 64.786 | 5.704 | 0.017* |
| | Mean unigram-to-bigram Delta P (tri_DP) | 46.269 | 0.508 | 0.476 |
| | Mean bigram-to-unigram Delta P (tri_2_DP) | 1.777 | 0.073 | 0.787 |
| | Proportion of 30,000 most frequent trigrams (tri_prop_30k) | 16.992 | 8.118 | 0.004** |

Reference group: intermediate proficiency, Significance level: * $p < 0.05$, ** $p < 0.01$

Of these four indices, *proportion of 30,000 most frequent trigrams* was identified as the most significant predictor, differentiating both basic-intermediate and intermediate-advanced proficiency groups. One likely explanation for this phenomenon lies in the nature of proportion indices that take into consideration n-grams that do not occur in the reference corpus. As noted by Bestgen and Granger (2014), n-grams that are absent in the reference corpus are either errors in learner language or creative combinations that are more likely to be used by advanced learners. They further observe that there is a negative correlation between the proportion of absent bigrams and the rated quality of English essays written by college-level L2 learners. While the results of the current study corroborate such findings, Crossley et al. (2012) in contrast report a weak but negative correlation between the proportion of n-grams that occur in the reference corpus and the holistic score of essays written by native-speaking college freshmen. They conclude that such findings support the position that essays of higher quality contain less frequent linguistic features. While the reason for this disparity is unclear, one possible explanation lies in the different range of proficiency levels captured by each study. That is, the proportion of n-grams that occur in the reference corpus could increase as learner errors decrease, but at a certain point begin to drop as writers use more novel combinations of their own.

The three other indices identified as significant predictors are measures of n-gram association strength. Their linear increase along with proficiency level indicates that writing of higher proficiency level contains more n-grams that are identified as strong collocations by native speakers. While this adds to the body of research on the relationship between association strength of n-grams and writing proficiency, the *mean bigram T* index showed a different pattern from that identified in previous studies. That is, both Durrant and Schmitt (2009) and Granger and Bestgen (2014) found that native speakers and more proficient writers tend to use less high-frequency bigrams identified by high *t*-scores but more strongly associated bigrams identified by high MI. While this inconsistency could partly be explained by the methodological difference from having used the TAALES (e.g., a different reference corpus, not considering part of speech) to calculate the indices of association strength, a more likely explanation could be provided by the relatively low proficiency level represented by the YELC, as mentioned above. Within this proficiency range, learners in the advanced group not only used more n-grams that simply occur in the reference corpus, but also used those that occur more frequently. This outcome seems natural in light of the short and error-prone nature of the essays in the basic-level group. From a pseudo-longitudinal perspective of learner corpus research that posits language use at different proficiency levels reflect the longitudinal development of language use, such a tendency in both the proportion and mean indices could be indicative of the developmental path of beginner-level Korean EFL learners' English writing ability.

Another noteworthy finding from the initial correlation analysis is that the Delta P indices represent a different perspective of n-gram use from other indices (i.e., they are not strongly correlated to other indices). However, the relative lack of attention to the Delta P index in the literature (see Ebeling & Hasselgård, 2015; Manning & Schütze, 1999) makes it difficult to interpret them qualitatively. Further research should be conducted in the future to examine which n-grams are emphasized by high Delta P indices, compared to other indices of association strength such as *t*-score and MI.

## 2. RQ 2: Predicting Proficiency Groups of Unseen Essays

The multinomial logistic regression model was applied to essays in both the training set and the test set to examine the extent to which indices based on n-grams are predictive of L2 writing proficiency. The regression model correctly classified 62.1% of essays in the training set and 52.5% of essays in the test set according to proficiency group. It demonstrated the highest accuracy for basic-level essays in the training set (73.1%), and advanced-level essays in the test set (68.2%). On the contrary, the regression model showed the lowest classification accuracy for intermediate-level essays in both the training set (47.6%) and the test set (28.9%). The fact that intermediate-level essays in the test set were accurately categorized at a below chance level (i.e., 33.3%) indicates that the eight independent variables were not able to capture the characteristics of the essays in the intermediate proficiency group. The classification results are summarized in Table 5.

**TABLE 5**

Classification results of the regression model

| Set | Predicted / Observed | Basic | | Intermediate | | Advanced | | Total | |
|---|---|---|---|---|---|---|---|---|---|
| | | *N* (%) | | *N* (%) | | *N* (%) | | *N* (%) | |
| Training set | Basic | 60 | (73.1) | 15 | (18.3) | 7 | (8.5) | 82 | (100.0) |
| | Intermediate | 25 | (30.5) | 39 | (47.6) | 18 | (22.0) | 82 | (100.0) |
| | Advanced | 8 | (10.5) | 18 | (23.7) | 50 | (65.8) | 76 | (100.0) |
| Test set | Basic | 22 | (57.9) | 7 | (18.4) | 9 | (23.7) | 38 | (100.0) |
| | Intermediate | 20 | (52.6) | 11 | (28.9) | 7 | (18.4) | 38 | (100.0) |
| | Advanced | 5 | (11.4) | 9 | (20.5) | 30 | (68.2) | 44 | (100.0) |

Overall accuracy on training set: 149/240 = 62.1%, On test set: 63/120 = 52.5%

While the classification results of the regression model showed an overall accuracy rate (52.5%) that is higher than the baseline expected by chance (i.e., 33.3%) on the test set, this accuracy rate is not very high compared to previous studies that have attempted to predict proficiency levels based on computational indices (Crossley et al., 2014; Kim, 2014; McNamara et al., 2010). This indicates that indices based on n-grams need to be complemented by other linguistic features in order to provide a better account of human judgment on writing proficiency. Knowing that n-grams simultaneously reflect lexical and syntactic features of the text (Crossley et al., 2012), it remains to be seen how indices based on n-grams can complement other significant predictors identified in the literature when predicting L2 writing proficiency.

## Ⅴ. Conclusion

The results of this study reveal how computational indices based on n-grams can be used to predict L2 writing proficiency. The multinomial logistic regression analysis identified four indices as significant predictors of proficiency groups: *mean bigram T*, *mean bigram Delta P*, *mean bigram-to-unigram Delta P*, and *proportion of 30,000 most frequent trigrams*. Furthermore, this regression model could to some extent classify unseen essays according to their proficiency groups. This lends support to the findings of previous studies that have shed light on the relationship between linguistic features and writing proficiency (Crossley & McNamara, 2012; Crossley et al., 2014; Jung et al., 2015; Kim, 2014; McNamara et al., 2010). In particular, the present research complements these studies by examining the relationship between indices based on n-grams and writing proficiency, which is beginning to receive attention in the literature (see Bestgen & Granger, 2014; Crossley et al., 2012).

Nonetheless, the present study is not without its limitations. The corpus used in this study was built based on data from learners at a single university, and therefore the results may only generalize to its closest peers such as other college-level Korean EFL learners. Additional research is required to determine the extent to which the findings are generalizable to writing from other grade levels (e.g., secondary education). Another concern with using proficiency data from the YELC is in its grading process. Although assigning a holistic proficiency level based on the results of the entire placement test is more valid compared to other extrinsic methods of operationalizing proficiency (e.g., age, length of study), the fact that writing proficiency level was not purely based on the quality of the argumentative essays in this study could undermine the findings. There were also limitations in terms of the amount of text from each learner. A preliminary analysis of the essays revealed that learners across all proficiency levels were copying sequences of words directly from the task prompt. Such prompt influence could have confounded the systematic difference in the indices across proficiency groups, especially given the little amount of text from each learner (i.e., an average of 254.96 words). Therefore, the results and implications of this study need to be evaluated incorporating these research limitations.

To mitigate the issue related to the grading process, a replication study could be conducted with the same data after adopting a rigorous post-hoc grading procedure (e.g., the ‘2+1 procedure’ in which two raters grade the essays and a third rater intervenes in case of severe disagreement) for a clearer picture regarding the relationship between indices based on n-grams and L2 writing proficiency. Furthermore, a qualitative analysis of n-grams in learner writing could be conducted to identify the characteristics of n-grams that influence these computational indices (e.g., n-grams that do not occur in the reference corpus). The findings of such studies could be useful for validating the indices for their potential future application to language assessment. Finally, these indices based on n-grams could be examined in conjunction with other computational indices (e.g., indices representing syntactic and lexical complexity) to identify any significant relationship amongst themselves in the context of predicting L2 writing proficiency.

## REFERENCES

Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 511-533). Cambridge: Cambridge University Press.

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.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgård & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Essex: Pearson Education Limited.

Chen, Y.-H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849-880.

Choe, J.-W., & Song, J.-Y. (2013). The topical classification of essays by college student English learners using hierarchical clustering. *Language Information*, *17*(5), 93-115.

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

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In G. M. Youngblood & P. M. McCarthy (Eds.), *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference* (pp. 214-219). Palo Alto, CA: The AAAI Press.

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

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157-177.

Ebeling, S. O., & Hasselgård, H. (2015). Learner corpora and phraseology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 207-229). Cambridge: Cambridge University Press.

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Doctoral dissertation, University of Stuttgart, Stuttgart.

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

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

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

Jung, Y., Crossley, S. A., & McNamara, D. S. (2015). *CaMLA working papers 2015-05: Linguistic features in MELAB writing task performances*. Ann Arbor, MI: CaMLA.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Kim, J.-Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, *69*(4), 27-51.

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

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190-208.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McDonald, J. H. (2014). *Handbook of biological statistics* (3rd ed.). Baltimore, MD: Sparky House Publishing.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57-86.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Rhee, S.-C., & Jung C. K. (2014). Compilation of the Yonsei English Learner Corpus

(YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *Journal of the Korea Contents Association*, *14*(11), 1019-1029.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*, 214-225.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Honolulu, HI: University of Hawaii Press.

## APPENDIX

### 1. Indices used in the current study

| Index Name | Description |
|---|---|
| Bigram_Frequency | Mean bigram frequency score |
| Bigram_Range | Mean bigram range score |
| Bigram_Frequency_Log | Mean bigram frequency score |
| Bigram_Range_Log | Mean bigram range score |
| bi_MI | Mean Mutual Information score |
| bi_MI2 | Mean Mutual Information score (MI^2) |
| bi_T | Mean T association strength score |
| bi_DP | Mean Delta P association score (left to right) |
| bi_AC | Mean Approximate Collexeme strength score |
| | (left to right DP * frequency of first item) |
| bi_prop_10k | Proportion of 10,000 most frequent bigrams |
| bi_prop_20k | Proportion of 20,000 most frequent bigrams |
| bi_prop_30k | Proportion of 30,000 most frequent bigrams |
| bi_prop_40k | Proportion of 40,000 most frequent bigrams |
| bi_prop_50k | Proportion of 50,000 most frequent bigrams |
| bi_prop_60k | Proportion of 60,000 most frequent bigrams |
| bi_prop_70k | Proportion of 70,000 most frequent bigrams |
| bi_prop_80k | Proportion of 80,000 most frequent bigrams |
| bi_prop_90k | Proportion of 90,000 most frequent bigrams |
| bi_prop_100k | Proportion of 100,000 most frequent bigrams |
| Trigram_Frequency | Mean trigram frequency score |

| | |
|---|---|
| Trigram_Range | Mean trigram range score |
| Trigram_Frequency_Log | Mean trigram frequency score |
| Trigram_Range_Log | Mean trigram range score |
| tri_MI | Mean Mutual Information score (unigram-to-bigram) |
| tri_MI2 | Mean Mutual Information score (MI^2) (unigram-to-bigram) |
| tri_T | Mean T association strength score (unigram-to-bigram) |
| tri_DP | Mean Delta P association score (left to right) (unigram-to-bigram) |
| tri_AC | Mean Approximate Collexeme strength score |
| | (left to right DP * frequency of first item, unigram-to-bigram) |
| tri_2_MI | Mean Mutual Information score (bigram-to-unigram) |
| tri_2_MI2 | Mean Mutual Information score (MI^2) (bigram-to-unigram) |
| tri_2_T | Mean T association strength score (bigram-to-unigram) |
| tri_2_DP | Mean Delta P association score (left to right) (bigram-to-unigram) |
| tri_2_AC | Mean Approximate Collexeme strength score |
| | (left to right DP * frequency of first item, bigram-to-unigram) |
| tri_prop_10k | Proportion of 10,000 most frequent trigrams |
| tri_prop_20k | Proportion of 20,000 most frequent trigrams |
| tri_prop_30k | Proportion of 30,000 most frequent trigrams |
| tri_prop_40k | Proportion of 40,000 most frequent trigrams |
| tri_prop_50k | Proportion of 50,000 most frequent trigrams |
| tri_prop_60k | Proportion of 60,000 most frequent trigrams |
| tri_prop_70k | Proportion of 70,000 most frequent trigrams |
| tri_prop_80k | Proportion of 80,000 most frequent trigrams |
| tri_prop_90k | Proportion of 90,000 most frequent trigrams |
| tri_prop_100k | Proportion of 100,000 most frequent trigrams |

Reference corpus: academic sub-corpus of COCA

## 2. ANOVA results: means (SD) and *F* value

| Index | Basic (*N*=82) | | Intermediate (*N*=82) | | Advanced (*N*=76) | | *F* |
|---|---|---|---|---|---|---|---|
| Bigram_Frequency | 146.629 | (101.580) | 177.039 | (111.566) | 179.991 | (70.249) | 3.085* |
| Bigram_Range | 0.121 | (0.028) | 0.133 | (0.028) | 0.137 | (0.018) | 10.087** |
| Bigram_Frequency_Log | 1.164 | (0.133) | 1.198 | (0.127) | 1.230 | (0.081) | 7.494** |
| Bigram_Range_Log | -1.458 | (0.112) | -1.430 | (0.103) | -1.405 | (0.065) | 6.846** |
| **bi_MI** | **1.515** | **(0.228)** | **1.584** | **(0.190)** | **1.665** | **(0.198)** | **9.809**** |
| bi_MI2 | 8.543 | (0.391) | 8.692 | (0.364) | 8.845 | (0.289) | 15.617** |
| **bi_T** | **35.656** | **(13.710)** | **40.978** | **(10.594)** | **42.385** | **(7.840)** | **7.300**** |
| **bi_DP** | **0.035** | **(0.010)** | **0.042** | **(0.011)** | **0.050** | **(0.011)** | **41.708**** |
| bi_AC | 8080.409 | (5763.645) | 9723.443 | (5971.461) | 9836.223 | (3876.369) | 2.734 |
| bi_prop_10k | 0.332 | (0.067) | 0.347 | (0.060) | 0.377 | (0.056) | 11.163** |
| bi_prop_20k | 0.391 | (0.071) | 0.413 | (0.072) | 0.447 | (0.063) | 14.012** |
| bi_prop_30k | 0.428 | (0.070) | 0.451 | (0.078) | 0.482 | (0.067) | 12.327** |
| bi_prop_40k | 0.455 | (0.068) | 0.475 | (0.079) | 0.509 | (0.069) | 12.427** |
| bi_prop_50k | 0.480 | (0.069) | 0.496 | (0.081) | 0.533 | (0.073) | 11.262** |
| bi_prop_60k | 0.501 | (0.070) | 0.519 | (0.080) | 0.554 | (0.072) | 11.244** |
| bi_prop_70k | 0.513 | (0.071) | 0.532 | (0.080) | 0.568 | (0.073) | 11.706** |
| bi_prop_80k | 0.529 | (0.070) | 0.550 | (0.078) | 0.584 | (0.069) | 12.608** |
| bi_prop_90k | 0.538 | (0.070) | 0.560 | (0.079) | 0.594 | (0.069) | 12.970** |
| bi_prop_100k | 0.548 | (0.071) | 0.569 | (0.080) | 0.605 | (0.068) | 13.618** |
| Trigram_Frequency | 8.461 | (4.442) | 10.572 | (6.520) | 10.051 | (3.814) | 4.109* |
| Trigram_Range | 0.023 | (0.010) | 0.027 | (0.013) | 0.026 | (0.008) | 4.082* |
| Trigram_Frequency_Log | 0.421 | (0.134) | 0.480 | (0.119) | 0.479 | (0.087) | 5.897** |
| Trigram_Range_Log | -2.079 | (0.133) | -2.023 | (0.112) | -2.023 | (0.084) | 5.677** |
| tri_MI | 2.677 | (0.418) | 2.772 | (0.337) | 2.813 | (0.272) | 2.964 |
| tri_MI2 | 8.001 | (0.617) | 8.236 | (0.475) | 8.271 | (0.379) | 5.793** |
| tri_T | 16.044 | (3.792) | 17.688 | (4.603) | 17.387 | (2.907) | 4.161* |
| **tri_DP** | **0.004** | **(0.002)** | **0.005** | **(0.003)** | **0.006** | **(0.003)** | **12.041**** |
| tri_AC | 595.743 | (337.206) | 750.442 | (485.220) | 711.274 | (286.850) | 3.845* |
| tri_2_MI | 2.620 | (0.439) | 2.777 | (0.336) | 2.798 | (0.293) | 4.848** |
| **tri_2_MI2** | **7.941** | **(0.582)** | **8.237** | **(0.474)** | **8.256** | **(0.377)** | **9.023**** |
| **tri_2_T** | **15.184** | **(5.304)** | **17.346** | **(4.427)** | **17.132** | **(2.882)** | **4.900**** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **tri_2_DP** | **0.127** | **(0.039)** | **0.153** | **(0.033)** | **0.164** | **(0.037)** | **19.603**** |
| tri_2_AC | 565.217 | (328.758) | 720.543 | (477.139) | 682.980 | (279.707) | 4.122* |
| tri_prop_10k | 0.062 | (0.030) | 0.080 | (0.028) | 0.092 | (0.026) | 23.069** |
| tri_prop_20k | 0.080 | (0.035) | 0.101 | (0.032) | 0.122 | (0.032) | 30.902** |
| **tri_prop_30k** | **0.092** | **(0.037)** | **0.114** | **(0.035)** | **0.138** | **(0.036)** | **32.307**** |
| tri_prop_40k | 0.104 | (0.041) | 0.125 | (0.038) | 0.153 | (0.040) | 28.294** |
| tri_prop_50k | 0.113 | (0.044) | 0.134 | (0.041) | 0.163 | (0.041) | 26.939** |
| tri_prop_60k | 0.119 | (0.045) | 0.140 | (0.042) | 0.171 | (0.043) | 27.931** |
| tri_prop_70k | 0.126 | (0.047) | 0.150 | (0.046) | 0.179 | (0.045) | 25.880** |
| tri_prop_80k | 0.131 | (0.048) | 0.155 | (0.048) | 0.186 | (0.047) | 26.117** |
| tri_prop_90k | 0.135 | (0.050) | 0.158 | (0.049) | 0.191 | (0.049) | 25.480** |
| tri_prop_100k | 0.139 | (0.051) | 0.164 | (0.051) | 0.197 | (0.049) | 26.181** |

Significance level: * $p < 0.05$, ** $p < 0.01$, Indices in bold included in the regression model

Oh, Byung-Doh
Dept. of English Language Education at Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul
Email: byungdoh@snu.ac.kr

Received on 31 October 2017
Reviewed on 5 December 2017
Revised version received on 16 December 2017
Accepted on 22 December 2017