THE JOURNAL OF ASIA TEFL
Vol. 16, No. 1, Spring 2019, 37-52
http://dx.doi.org/10.18823/asiatefl.2019.16.1.3.37

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Predicting Second Language Writing Proficiency in Learner Texts Using Computational Tools*

YeonJoo Jung
Pusan National University

Scott Crossley
Georgia State University

Danielle McNamara
Arizona State University

This study explores whether linguistic features can predict second language writing proficiency in the Michigan English Language Assessment Battery (MELAB) writing tasks. Advanced computational tools were used to automatically assess linguistic features related to lexical sophistication, syntactic complexity, cohesion, and text structure of writing samples graded by expert raters. The findings of this study show that an analysis of linguistic features can be used to significantly predict human judgments of the essays for the MELAB writing tasks. Furthermore, the findings indicate the relative contribution of a range of linguistic features in MELAB essays to overall second language (L2) writing proficiency scores. For instance, linguistic features associated with text length and lexical sophistication were found to be more predictive of writing quality in MELAB than those associated with cohesion and syntactic complexity. This study has important implications for defining writing proficiency at different levels of achievement in L2 academic writing as well as improving the current MELAB rating scale and rater training practices. Directions for future research are also discussed.

Keywords: linguistic features, second language writing proficiency, computational analysis

Introduction

In the past few decades, a considerable amount of research has been conducted to elucidate the distinct nature of second language (L2) writing and the development of L2 writing competencies (e.g., Friginal, Li, & Weigle, 2014; Grant & Ginther, 2000; Lee, Bychkovska, & Maxwell, 2018). The majority of this research has shown the potential for linguistic features to characterize L2 writing by describing salient linguistic differences between first language (L1) and L2 or by illustrating linguistic development in L2 writers. In an overview of 72 empirical investigations that compared L1 and L2 writing, Silva (1993) presented numerous notable features of written texts produced in L2 English. These features include shorter T-units, fewer but longer clauses, more coordination, less subordination, less noun modification, less passivization, and fewer cohesion devices including more conjunctives and lexical ties. In addition,

* An earlier version of this paper was published in CaMLA Working Papers 2015-05.

L2 writers were likely to demonstrate less lexical control, variety, and overall lexical sophistication.

Previous studies have also illuminated how writing competencies develop as L2 writers become more proficient. Among different aspects of writing competencies that learners need to acquire to be able to write effectively in their L2, mastery of linguistic competencies has merited substantial attention in research (Barkaoui, 2007). Such research has investigated L2 writing in terms of the orthography, morphology, lexicon, syntax as well as the discourse and rhetorical conventions of L2. For example, L2 writers are expected to attain the ability to produce lengthy texts which contain a broad range of sophisticated vocabulary, appropriate metadiscourse features such as connectives and hedges, and complex syntactic structures (e.g., Buckwalter & Lo, 2002) as well as the ability to employ appropriate patterns of text organization (e.g., descriptive, persuasive, narrative). Furthermore, researchers have emphasized the importance of linguistic aspects in L2 writing, suggesting that proper use of linguistic features is required for L2 writers to develop writing strategies such as planning, drafting, and revising (e.g., Chenoweth & Hayes, 2001).

With a particular focus on linguistic features in L2 texts written by learners at different proficiency levels, past studies found that more proficient L2 writers tend to use more words with more letters or syllables in their essays (Grant & Ginther, 2000; Reppen, 1994). With regard to syntax, higher-rated L2 essays contain more surface code measures such as subordination (Grant & Ginther, 2000), instances of passive voice (Connor, 1990; Grant & Ginther, 2000), instances of nominalisations, prepositions (Connor, 1990), pronouns (Reid, 1992) and fewer present tense forms (Reppen, 1994). Furthermore, researchers have shown that the more proficient L2 writers become, the more they produce explicit cohesion devices (i.e., connectives) (Connor, 1990).

Recent advances in various disciplines such as computational linguistics, discourse processing, and information retrieval have made it possible to computationally investigate measures of cohesion, rhetorical choices, and linguistic sophistication in the assessment of writing proficiency. Together, these advances allow for the automated analysis of many surface and deep level factors, such as lexical sophistication, syntactic complexity, text cohesion, rhetorical features, and text structure, allowing detailed analyses of language to take place. Such systems are known as Automatic Essays Scoring (AES) systems.

These AES systems, however, are not without limitations. For instance, a recent study conducted by Crossley, Kyle, Allen, Gou, and McNamara (2014) demonstrated some of the weaknesses of AES systems. These weaknesses included a limited portion of writing construct and the restricted writing genres that can be assessed by AES systems. More specifically, AES systems cannot assess the entire construct of writing because they fail to address issues of argumentation, purpose, audience, and rhetorical effectiveness, which are hallmarks of quality writing attended to by human raters (e.g., Condon, 2013). Moreover, AES systems are generally only successful at assessing limited writing genres such as the independent writing genre and less successful at assessing other genres such as authentic performance tasks and portfolio based-writing, which are considered more credible and valid forms of writing (e.g., Condon, 2013).

An increasing volume of research has used computation tools to examine writing proficiency. The tools most widely used have likely been Coh-Metrix, which is freely available to researchers (McNamara, Graesser, McCarthy, & Cai, 2014). Coh-Metrix indices have been used to explore relationships between essay scores assigned in standardized writing tests and linguistic features (Crossley & McNamara, 2012; Guo, Crossley, & McNamara, 2013). Studies using Coh-Metrix have reported similar findings in terms of the linguistic features identified as significant in human judgments of essays to those using e-rater. That is to say, lexical sophistication variables included in e-rater were revealed to be integral to predicting human judgments of essay quality in past studies (e.g., Enright & Quinlan, 2010), as was also shown in the following studies that used Coh-Metrix. For example, in Guo et al. (2013) linguistic variables related to lexical sophistication, syntactic complexity, and, to some degree, cohesion (e.g., noun hypernymy values, past participle verbs, conditional connectives) were found to contribute to the prediction of human scoring of TOEFL independent writing tasks. The study also helped validate the use of scoring rubrics by

verifying that some of the predictors are meaningfully related to the writing aspects specified in the rubrics. Furthermore, Guo et al. suggested that the rubrics were in need of revision by showing that features that are not specified in the rubrics were found to be attended to by raters when deriving judgments of writing quality. A more recent study conducted by Crossley et al. (2014) used indices reported by Coh-Metrix to investigate the potential for linguistic features to predict L2 writing proficiency in TOEFL writing samples. They showed that linguistic features related to breadth of lexical production, lexical sophistication, key words use, local and global cohesion, and tense are important predictors of L2 writing quality.

Despite its contributions, concerns have been raised regarding Coh-Metrix. Foremost, the tool is over ten years old and many new potential features have not been incorporated. These features include indices of global cohesion (i.e., cohesion between larger chunks of language) and newer indices of lexical sophistication. While Coh-Metrix may not include these indices, new tools such as TAALES (Kyle & Crossley, 2015) and TAACO (Crossley, Kyle, & McNamara, 2016) do. These tools, like Coh-Metrix, have been validated in a number of studies (Crossley et al., 2016; Kyle & Crossley, 2015).

As presented thus far, there is mounting evidence supporting the validity of computational tools to contribute to our knowledge of L2 writing. More specifically, a large volume of studies that examined L2 writing development using computational tools have demonstrated the strength of automated indices, which were selected based on face validity by researchers, to predict human evaluations of L2 writing products. They also suggested that the predictive models established through linear regression analyses can be applied to understand the construct of L2 writing proficiency. Nonetheless, it may be premature to tout the strength of the indices in measuring L2 writing proficiency because, while the indices have been validated across numerous studies (e.g., Crossley & McNamara, 2012; Guo et al., 2013; Kyle & Crossley, 2015), the majority of such studies have focused on the writing portion of the TOEFL, and thus the generalizability of the reported findings outside the context of a particular application has yet to be addressed. Moreover, as Messick (1989) suggested, evidence of validity should be continuously gathered across different contexts. As such, more research is called for to ensure the validity of computational indices for evaluating L2 essay quality in different large-scale assessment contexts other than TOEFL. The current study uses three computer programs (Coh-Metrix, TAALES, & TAACO) to assess writing quality in a similar large-scale language test to TOEFL, the Michigan English Language Assessment Battery (MELAB). The results of this study shed light on not only the linguistic characteristics of writing that are used to distinguish levels of L2 writing proficiency on the MELAB, but also the validity of the indices collected from the automated tools for assessing L2 writing proficiency in general. The results will also help advance our understanding of the construct of L2 writing proficiency. Furthermore, as indicated by Weigle (2013), research using advanced computational tools can provide useful information for rating scale development and rater training as well as the improvement of automated scoring algorithms. With particular regard to the MELAB writing task, results of such research would facilitate the interpretations of learners' writing competence represented as a human-assigned holistic score and help validate the rating scale used by raters to assess learner essays. This study was guided by the following research question:

What linguistic features, as measured by Coh-Metrix, distinguish MELAB test-taker writing performance as represented as a single summary score awarded by expert raters on the basis of the ten-level MELAB Composition Rating Scale?

Method

In the current study, we used Coh-Metrix, TAALES, and TAACO to investigate whether and how linguistic features related to four variable classes (i.e., text structure, cohesion, lexical sophistication, and syntactic complexity) help characterize L2 writing proficiency in the MELAB writing test. To accomplish the goal of better understanding the relationships between human judgments of writing quality and the

language features that differ as a function of these judgments, we assessed language performance differences in a corpus of MELAB written text samples using statistical analysis and machine learning techniques.

MELAB Essay Corpus

A corpus of 750 essays was collected from MELAB writing tests administered in 2013. The writing samples were stratified according to score level, gender, and age in order to ensure that the findings of the study could be generalized to the entire MELAB test-taking population. The writing samples came from 21 test forms, each of which included a differing set of prompt choices. Hence, the essay samples in the current dataset were responses to 42 different prompts. The essays, test-takers' demographic information (i.e., L1, age, gender), the final composition scores, and the scores for other sections (i.e., listening, grammar cloze vocabulary reading [GCVR], speaking) were provided directly by Cambridge Michigan Language Assessments (CaMLA). The 750 test-takers represent 62 different L1 backgrounds (two test takers did not indicate their L1s). Because the MELAB essays were hand-written, they were transcribed and saved into .txt files for subsequent computational analyses.

MELAB Writing Task

The time limit for the composition was 30 minutes. Each test-taker was given a test booklet with instructions and two topics to choose from. The test-takers were asked to choose one of the topics and write at least a one-page long (200-300 words) composition about the selected topic. The instructions indicated that a composition containing less than 150 words would be likely given a lower score.

Human Ratings of Writing Proficiency

Each essay is rated independently by at least two trained raters using locally developed standardized holistic rubrics. The MELAB composition rating scale has ten levels, ranging from 53 to 97, with nearly equal intervals (i.e., 97, 93, 87, … 57, 53). Because midpoints between the levels are also possible, there are a total of 19 potential scores, and each essay is assigned one of these scores. If the human rater scores differ by one score point, the average of the two holistic scores is used as the final composition score. If the human raters give non-adjacent ratings, the essay is read by a third, experienced rater. A successful essay with a score of 97 is described as:

> Topic is richly and fully developed. Flexible use of a wide range of syntactic (sentence-level) structures, accurate morphological (word forms) control. Organization is appropriate and effective, and there is excellent control of connection. There is a wide range of appropriately used vocabulary. Spelling and punctuation appear error free.

On the other hand, the lowest score (53) is assigned to an essay when the essay is:

> Extremely short, usually about 40 words or less; communicates nothing, and is often copied directly from the prompt. There is little sign of syntactic or morphological control, and no apparent organization or connection. Vocabulary is extremely restricted and repetitively used. Spelling is often indecipherable and punctuation is missing or appears random.

As seen above, the MELAB writing rating scale appears comprehensive, yet not exhaustive. More specifically, the descriptors of the rating scale address various aspects of writing such as topic development, text length, breadth and appropriateness of lexical choice, adequate morphological and syntactic control, cohesion, and accuracy of spelling and punctuation. However, the rubric does not

provide specific guidance to the raters with regard to which specific features they are expected to attend to in assessing each of those aspects. For example, the rating scale does not specify how cohesion should be achieved in an essay (e.g., through cohesive devices such as causality, connectives, lexical overlap). The descriptors might be explained in detail and fully exemplified during the rater training, but such information is not reported in any publicly available MELAB documents.

The inter-rater reliability for the 2013 MELAB writing test dataset is available in the MELAB 2013 Report (CaMLA, 2013). CaMLA monitors the mean exact and adjacent agreement between the first and second rater and the mean Pearson correlation of the scores awarded by each rater. In 2013 these were 90.03% and 0.83 respectively.

Variable Selection

Coh-Metrix reports on over 800 variables in more than 11 categories that measure many surface and deep level factors, such as lexical sophistication, syntactic complexity, text cohesion, rhetorical features, and text structure, allowing detailed analyses of language to take place. TAALES measures 135 indices of lexical sophistication (e.g., word frequency, range, psycholinguistic word information), whereas TAACO assesses cohesion by calculating text scores for 150 indices related to lexical overlap, type-token ratio (TTR), givenness, and connectives (e.g., logical, positive, logical intentional, causal).

In this study, text length, lexical sophistication, cohesion, and syntactic complexity were chosen as construct relevant categories to assess MELAB essay quality. These categories have been shown to correlate with L2 essay scores, correspond to aspects of L2 writing which human raters attend to (e.g., Grant & Ginther, 2000), and are also features that are meaningfully related to aspects of writing specified in the MELAB composition rating scale. Hence, indices associated with these categories were used in the current study so as to discover the relationship between human judgments of writing quality and linguistic features that differ as a function of essay scores. The selected variables are discussed briefly below. We refer to the reader to Coh-Metrix (McNamara et al., 2014), TAALES (Kyle & Crossley, 2015), and TAACO (Crossley et al., 2016) for additional information.

Text structure

Coh-Metrix reports basic textual information for the number of words and the number of paragraphs. The number of words is calculated using the output from the Charniak parser, while the number of paragraphs is delimited by a hard return.

Lexical sophistication. Lexical sophistication is evaluated by both Coh-Metrix on the basis of syllables per word, word hypernymy and polysemy values, word frequency, psycholinguistic word properties, and lexical diversity. TAALES measures for lexical sophistication include word frequency, range, and the psycholinguistic properties of words. Hypernymy and polysemy values are calculated for all words in a given text have entries in the WordNet database (Fellbaum, 1998). Word and n-gram frequency is calculated from the CELEX corpus (Baayen, Piepenbrock, & van Rijn, 1993), and the written and spoken sections of the British National Corpus (BNC; 2007; Brown, 1984). Frequency indices are calculated for content words, function words, and all words. A less commonly used but still important measure of frequency is range indices (Kyle & Crossley, 2015). TAALES range indices account for how widely a word or word family is used, usually by providing a count of the number documents in which that word occurs using the BNC.

Both Coh-Metrix and TAALES use human ratings provided by the Medical Research Council Psycholinguistic Database (MRC; Wilson, 1988) to report the psycholinguistic properties for familiarity, concreteness, imageability, and meaningfulness. In addition, TAALES evaluates age of acquisition based on Kuperman, Stadthagen-Gonzales, and Brysbaert's (2012) AoA (Age of Acquisition) list and

concreteness based on Brysbaert, Warriner, and Kuperman (2014). Concreteness is rated based on perceptions of how simple it is to describe the meaning of a word. Concrete words are more tangible than abstract words, while familiar words are more readily recognized. Imageability scores indicate how easily a word can evoke a mental image and meaningfulness relates to the number of associations a word has with other words (Toglia & Battig, 1978). Age of acquisition (AoA) ratings are based on human judgments of the age at which a particular word is learned. These properties have successfully explained holistic scores of writing quality (e.g., Crossley & McNamara, 2012).

Syntactic complexity

Coh-Metrix estimates syntactic complexity using indices related to the number of higher level constituents per word, number of words before the main verb, and the number of modifiers per noun phrase, and syntactic categories for words (i.e., part of speech [POS] tags). Coh-Metrix also generates frequency data for the major syntactic categories adopted from the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993) and the Brill (1995) part of speech (POS) tagger. The POS categories are divided into content (e.g., noun, verb, adjective, adverb) and function words (e.g., preposition, pronoun, demonstratives). The POS categories also include syntactic information such as incidence of infinitive forms and embedded clauses. At the phrasal level, Coh-Metrix calculates the incidence of noun, verb, and prepositional phrases.

Cohesion

Cohesion refers to the presence or absence of the linguistic markers in the text that cue the reader on how to make connections between the ideas in the text. Cohesion can be further divided into local and global cohesion. Local cohesion concerns the interrelatedness between sentences, while global cohesion concerns the interrelatedness between larger segments of discourse (i.e., paragraphs) (Louwerse & Graesser, 2005).

Coh-Metrix and TAACO measures cohesion by evaluating lexical overlap, causality, connectives, and semantic similarity. Coh-Metrix reports four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap. Coh-Metrix measures causality by calculating the ratio of causal particles (e.g., *after all, because, hence, if, though*) to causal verbs, number of causal verbs, and LSA overlap between words. Connective indices reported by Coh-Metrix include the number of all connectives and number of positive causal connectives (e.g., *because*) which can discriminate between high and low cohesion texts. Coh-Metrix estimates semantic similarity between text segments (i.e., sentences and paragraphs) by using different types of Latent Semantic Analysis measures (LSA; Landauer, Foltz, & Laham, 1998), which allow for the measurement of conceptual similarity between adjacent sentences or paragraphs. At the paragraph level, LSA similarity is measured between adjacent paragraphs and between introduction and body paragraphs, body and conclusion paragraphs, and introduction and conclusion paragraphs. TAACO calculates a number of paragraph overlap indices to assess global cohesion. These indices compute lemma overlap between two adjacent paragraphs and between three adjacent paragraphs using features such as average and binary lemma overlap, content word lemma overlap, function word lemma overlap, and lemma overlap for nouns, verbs, and pronouns. In addition, semantic overlap between paragraphs for nouns and for verbs is also calculated. TAACO indices measure local cohesion by calculating a number of sentence overlap indices.

Statistical Analysis

For the current study, 1,051 indices were initially selected from Coh-Metrix, TAALES, and TAACO, all of which are relevant to the four categories of writing aspects. We first checked whether the indices were normally distributed. Any indices that were found to lack normal distributions were removed from

further consideration (n = 383). The essay corpus was randomly divided into two sets based on a 67/33 split (Whitten, Frank, & Hall, 2011): a training set (n = 509) and a test set (n = 241). The two sets of data (i.e., training and test) were not stratified proportionally to the original corpus. However, given that random sampling has been suggested as the most rigorous form of sampling and since the sample size was large, we did not expect any potential problem in the method that we employed to divide the corpus. This method allowed the prediction of the performance of the constructed regression model on an independent corpus (the test set). Furthermore, use of both training and test sets enabled investigating whether the results could be generalizable to an entire population beyond the population randomly sampled.

Using the training set, correlations were run to determine whether there was a significant ($p < .05$) and meaningful ($|r| > .1$) relation between the indices and the essay scores. Of the 383 indices that demonstrated normal distributions, 88 indices showed significant and meaningful relations. All significant indices were then checked for multicollinearity to ensure none of those indices correlated at $|r| \geq .90$. If two or more indices were found to correlate at this rate with each other, only the index with the higher correlation with the essay scores was retained for the subsequent regression analysis. After this step, 77 indices remained. As a next step, a stepwise multiple regression analysis was conducted to examine which of the selected indices (predictor variables) were predictive of human essay ratings (a dependent variable) for the 509 essays in the training set. The yielded regression model was then applied to the remaining 241 essays in the test set to predict future performance on an independent dataset.

Results

Correlations: Training Set

Correlations were conducted between the indices and the human scores for the 509 essays in the training set. Eighty-eight indices demonstrated significant correlations with the holistic scores, but twenty-two of them were multicollinear with each other at $|r| \geq .90$. Therefore, eleven indices from each multicollinear pair that demonstrated higher correlation with the essay scores were retained. In total, 77 indices were entered into a stepwise multiple regression in order to determine whether the selected indices could predict the variance in the holistic MELAB essay scores (see Appendix).

Regression Analysis: Training Set

A stepwise regression analysis using the 77 indices as the predictor variables to predict the human scores yielded a significant model, $F(12, 496) = 56.249$, $p < .001$ $r = .759$, $r^2 = .576$. Twelve indices from Coh-Metrix, TAALES, and TAACO were found to be significant predictors of the essay scores: *number of words per text*, the logarithm-transformed normed frequency of normed trigrams based on the written portion of the BNC (*BNC Written Trigram Freq Normed [word] Log*), the logarithm-transformed frequency of content words based on the written portion of the BNC (*BNC Written Freq CW Log*), word familiarity of content words, word concreteness of all words (both unigrams and bigrams) based on Brysbaert (*Brysbaert Concreteness Combined AW*), LSA overlap between introduction and conclusion (*semantic similarity [LSA introduction to conclusion]*), part of speech (POS): to-infinitive, number of positive causal connectives, number of sentences, number of demonstratives, word frequency of content words based on CELEX (*CELEX word frequency [content words]*), and incidence of noun phrase. This model demonstrated that the twelve indices together explained 57.6% of the variance in the assessment of the 509 essays in the training set. These results are displayed in Table 1 and 2.

TABLE 1
The Final Model of the Stepwise Regression Analysis: Training Set

Entry	Variable added	Variable removed	R	R^2
1	Number of words per text		.618	.381
2	Kuperman AoA CW Log		.667	.445
3	BNC Written Trigram Freq Normed (word) Log		.701	.491
4	BNC Written Freq CW Log		.718	.516
5		Kuperman AoA CW Log	.717	.515
6	MRC Word familiarity (content words)		.726	.527
7	Brysbaert Concreteness Combined AW		.732	.535
8	Semantic similarity (LSA introduction to conclusion)		.737	.544
9	Part of speech: to-infinitive		.740	.548
10	Number of positive causal connectives		.745	.555
11	Number of sentences		.749	.561
12	Number of demonstratives		.753	.567
13	CELEX Word frequency (content words)		.757	.573
14	Incidence of noun phrase		.759	.576

TABLE 2
Coefficients of the Final Model: Training Set

Variable	Category	*B*	B	S.E.
Number of words per text	Text structure	.051	.207	.012
Kuperman AoA CW Log	Lexical sophistication			
BNC Written Trigram Freq Normed (word) Log	Lexical sophistication	25.151	.222	4.087
BNC Written Freq CW Log	Lexical sophistication	-7.456	-.207	1.824
MRC Word familiarity (content words)	Lexical sophistication	-.135	-.102	.061
Brysbaert Concreteness Combined AW	Lexical sophistication	-11.382	-.149	2.813
Semantic similarity (LSA introduction to conclusion)	Cohesion	3.314	.072	1.429
Part of speech: to-infinitive	Syntactic complexity	.071	.119	.020
Number of positive causal connectives	Cohesion	-48.223	-.156	12.196
Number of sentences	Text structure	.226	.144	.065
Number of demonstratives	Cohesion	65.498	.094	24.591
CELEX Word frequency (content words)	Lexical sophistication	-6.683	.140	2.725
Incidence of noun phrase	Syntactic complexity	-.026	-.077	.012

Regression Analysis: Test Set

The model derived from the training set was used to predict the human scores in the test set. To determine the predictive power of the twelve indices retained in the regression model, an estimated score for each essay in the test set was computed using the B weights and the constant from the training set regression analysis. This computation provided a score estimate for the essays in the test set. A Pearson's correlation was then conducted between the estimated score and the actual score assigned to each of the essays in the test set. This correlation, together with its r^2, was then calculated to determine the predictive accuracy of the training set regression model on the independent dataset. The regression model, when applied to the test set, reported r = .740, r^2= .548. The results from the test set model demonstrated that the combination of the twelve indices accounted for 54.8% of the variance in the assigned scores of the 241 essays in the test set, supporting the generalizability of the model.

Discussion

This study builds upon previous research that demonstrated the validity of AES systems in general and the use computational programs for assessing L2 writing proficiency. The purpose of the present study was to explore whether a wide range of automated indices produced by a number of tools (Coh-Metrix, TAALES, and TAACO) correspond to the construct of L2 writing proficiency that they are meant to measure in large-scale assessment (i.e., the MELAB) other than the TOEFL. To do so we examined the relationships between human judgments of essay quality and language features that differ as a function of these judgments in a corpus of MELAB independent essays. We also looked at whether our findings converge on or diverge from those of past studies.

Correlations between the computational indices and the MELAB writing scores demonstrated small to large effect sizes, providing a degree of construct validity for most of the indices selected and examined in this study. The 12 indices retained in the regression were found to cover the categories of text length, lexical sophistication, cohesion, and syntactic complexity. These indices were successful in predicting the holistic scores for the writing quality of the MELAB essays sampled. A multiple regression predicting holistic judgments of test-takers' writing proficiency using these automated indices explained 57.6% of the variance in a training set and 54.8% of the variance in a test set. The results also demonstrate that the models established from the training set are extendable to the test set (an independent dataset), showing similar predictive accuracy achieved in the training and test sets. This, in turn, demonstrates the strength of the model to predict human evaluations of essays in the test set. Thus, the results of the study lend reliable support to the notion that linguistic features are significant predictors of human scores for writing quality. The predictive ability of the yielded model is discussed in a greater detail below.

Significant Predictors of L2 Writing Proficiency

Similar to previous studies on L2 writing assessment in language test contexts (i.e., TOEFL), this study found that two indices related to textual structure have a large effect on the essay scores assigned by human raters, in that longer essays were rated higher. *Number of words per text* was found to be the strongest predictor of essay quality among the twelve indices that were retained in the regression model, accounting for 38.1% of the variance of the human judgments of essay quality. Also, the addition of *number of sentences* to the previous model increased the percentage of variance explained by the model by 0.6%. Although text length alone cannot be considered as signifying quality writing, many of the features of highly scored essays (e.g., substantial supporting details and idea development) are difficult to accomplish in a shorter essay (Chodorow & Burstein, 2004). In order to present a writer's intended message in a text, a certain amount of development is necessary by means of a minimum number of words.

For instance, in their study of the TOEFL independent essays, Enright and Quinlan (2010) found that measures of rhetorical control such as organization and development, which are highly correlated with text length, comprised 60% of the essay scores. Based on their findings, they suggested that despite the varying opinions as to the construct relevance of text length, it is a reliable proxy of writing quality in writing assessment, whether scored automatically or by humans. They also stated that text length can be considered as a construct relevant feature especially in a standardized test setting, where test takers are asked to carry out an impromptu essay task under a strict time limit. In this setting, the length of an essay may reflect fluent production, development, and elaboration. Guo et al. (2013) showed that the Coh-Metrix index for number of words per text accounted for 26.4% of the variance of the scores in TOEFL integrated essays and 47.8% in TOEFL independent essays. More recently, Crossley et al. (2014) also demonstrated that indices related to text length may indicate a learner's ability to properly organize and develop an essay even though essay length may not always be synonymous with essay quality for human raters.

The next strongest predictors were indices related to lexical sophistication: the logarithm-transformed

normed frequency of normed trigrams based on the written portion of the BNC (*BNC Written Trigram Freq Normed [word] Log*), the logarithm-transformed frequency of content words based on the written portion of the BNC (*BNC Written Freq CW Log*), word familiarity of content words, word concreteness of all words (both unigrams and bigrams) based on Brysbaert (*Brysbaert Concreteness Combined AW*), word frequency of content words based on CELEX (*CELEX word frequency [content words]*). The addition of these five indices to the previous model increased the percentage of variance of holistic scores explained by the model by 16 %. This is consistent with findings of previous studies, which have demonstrated that lexical indices related to breadth of lexical knowledge (e.g., frequency) and access to core lexical items (e.g., concreteness, familiarity) were highly predictive of general language proficiency in individual L2 learners (Crossley, Salsbury & McNamara, 2012). More specifically, the addition of the normed frequency of trigrams (based on the written portion of the BNC) to the previous model increased the percentage of variance explained by 4.6 % in the essay scores. This index was positively correlated with the human judgments, suggesting that essays containing more frequent trigrams were judged by the CaMLA raters to be more proficient than those with lower frequency trigrams. Past research on human scoring of L2 writing (e.g., Barkaoui, 2007) has suggested that raters attend to a number of aspects of writing, not all of which can be captured in a scoring rubric (e.g., the normed frequency of trigrams in this study). A follow-up study might consider addressing this issue using qualitative research methods such as think-aloud protocols and interviews.

The other two-word frequency indices, *BNC Written Freq CW Log* and *CELEX word frequency (content words)*, were also found to be predictive of quality writing in the MELAB, increasing the percentage of variance explained by the model by 2.5% and 0.6%, respectively. This shows that MELAB essays with higher scores tended to include fewer content words that are frequent in the BNC written corpus and the CELEX corpus. Finally, the MRC familiarity for content words and the Brysbaert concreteness for all words (combined) were also significant predictors in the model, each adding the percentage of variance explained by the model by 1.2% and 0.8% respectively. Both of these indices were negatively correlated with the essay scores, indicating that high-proficiency MELAB essays tended to contain less familiar content words and more abstract words than lower proficiency essays. These findings lend support to the previous research (e.g., Crossley, Roscoe, & McNamara, 2013; Kyle & Crossley, 2015) which demonstrated that the indices of frequency, psycholinguistic word information, and n-grams are valid measures of written lexical proficiency. However, the descriptors of the rating scale do not explicitly elaborate on lexical sophistication (see Method section above) and instead focus only on the breadth and appropriateness of lexical choices. This finding, therefore, may be a useful piece of information that could advise raters of what they may particularly attend to when evaluating the lexical sophistication of essays.

With regard to cohesion, an index of global cohesion (*semantic similarity [LSA introduction to conclusion]*), one local cohesive index (*number of positive causal connectives [e.g., arise, arising]*), and an index of givenness (*number of demonstratives*[1] *[e.g., this, that, these])*, which incorporates both global and local aspects of cohesion, were also found to be significant predictors of essay quality in the MELAB. The addition of these indices to the previous model increased the percentage of variance explained by 2.2 %. The semantic similarity score was positively correlated with the essay scores, demonstrating that conceptual similarity between the introductory and concluding paragraphs is a strong indicator of good writing quality. This finding may indicate that maintaining global cohesion throughout an essay is an important component of quality writing. Analyses of the local cohesion and givenness indices suggested that high proficiency writing samples included a smaller incidence of positive causal connectives (e.g., *as a consequence, as a result*) and a greater incidence of demonstratives. In other words, highly proficient writers tended to use fewer positive causal connectives and more demonstratives than low proficient writers to link ideas across clauses and sentences in an essay. This analysis of textual cohesion lends support to suggestions made regarding global cohesion and local cohesive indices and human judgments of writing proficiency in other studies. In other words, while the incidence of global cohesive devices has

[1] Demonstratives may be pronominal as well as determiners (i.e., *that* hat is nice vs. *that* is a nice hat).

been positively correlated with human judgments of writing proficiency, local cohesive devices have not been positively correlated to expert raters' judgments of essay quality, especially for independent essays (e.g., Guo et al., 2013). The findings of the current study support this notion by showing that when assessing the MELAB writing samples, positive causal connectives were negatively correlated with essay quality, but semantic similarity between paragraphs and demonstratives were positively correlated with essay quality.

This study revealed that two indices of syntactic complexity (*Part of speech: to-infinitive* and *Incidence of noun phrases*) were significant predictors of L2 writing proficiency. The addition of these two indices to the previous model increased the percentage of variance explained by the model by 0.4% and 0.3%, respectively. Analysis of both indices indicated that more syntactically complex sentences were related to higher scores. More specifically, the results indicate that as the syntax in an essay becomes complex through embedding more infinitival clauses, essay scores increase suggesting that L2 writers judged to be more advanced produced texts with more complex syntax in comparison to those who were judged to be less proficient, at least in the context of the MELAB. In addition, higher scores were assigned to essays with a higher incidence of noun phrases likely because such essays are structurally more complex and dense in terms of information contained. A higher incidence for noun phrase as well as other content words is likely an important index of how much substantive content a given text contains.

Implications for Assessing Writing Performance in MELAB

Overall, the current analysis demonstrates that linguistic features vary with the essay scores in MELAB writing performance. As described thus far, writing proficiency was partially determined by text length, the level of lexical and syntactic sophistication, and patterns of cohesive device use, all of which are not only aspects of writing specified in the MELAB composition rating scale but also the attributes that have been found to constitute the construct of L2 writing proficiency in previous research (e.g., Guo et al., 2013). Hence, this finding could be considered as compelling evidence that the automated indices correspond to the multiple aspects of the construct they are expected to measure and that the trained raters have appropriately used the scoring rubric in their evaluations of the MELAB essays.

However, some discrepancies were found between the aspects of writing described in the MELAB rating scales and the linguistic features that predicted the scores of learner essays. First of all, lexical sophistication indices were found to be highly predictive of the essay scores although the rating scale does not specify what aspects of vocabulary knowledge (i.e., word frequency, concreteness, familiarity) are deemed as important in scoring. While the rubric lists the breadth and appropriateness of lexical choice as one of many evaluative criteria, the rubric does not discuss depth of lexical knowledge an important aspect to attend to in scoring. Our regression analysis, however, demonstrated that indices related to depth of lexical knowledge (word familiarity and concreteness) were significant predictors of writing proficiency. Furthermore, production of multi-word units such as trigrams is not listed as one of the evaluative criteria, while the findings of this study suggest that an index of trigram frequency was an important indicator of lexical proficiency of the MELAB test-takers. The rating scale could be updated with guidance on how to assess the test takers' ability to properly use word combinations (multi-word sequences) to deliver their intended message.

Additionally, although the rubric emphasizes a test-taker's ability to properly use cohesive devices to create coherent discourse, our regression analysis showed that the more proficient writers used fewer local cohesive devices (positive causal connectives) and more global cohesion devices. One possible explanation for this may be *a reverse cohesion effect*. As also revealed in previous studies (e.g., Crossley et al., 2015), high knowledge readers, such as the expert raters used in this study seem to be influenced in their judgments of essay coherence and quality not from local cohesion, but from global cohesion that connects ideas across larger segments of the text.

Conclusion

This study demonstrated that an analysis of linguistic features can predict human judgments of the essays for the MELAB writing tasks. In addition, the findings elucidated the relative contribution of a range of linguistic features in MELAB essays to overall L2 writing proficiency scores. In this study, linguistic features associated with text length and lexical sophistication were revealed to have greater predictive power for writing quality in MELAB than those associated with cohesion and syntactic complexity. Given that this study employed rigorous statistical methodology (training and test sets), which allowed us to investigate the generalizability of the model, we have confidence that the current findings are extendable to the MELAB test-taker population. Although it may be premature to generalize these findings beyond the MELAB, the linguistic aspects of test performance identified in this study and previous research may have the potential to provide a sketch of the features that may distinguish L2 learners' writing abilities at various proficiency levels.

Overall, the findings of the present study provide support for the use of computational indices to examine human evaluations of L2 writing proficiency. Moreover, the current findings demonstrate that the computational indices corresponding to the four aspects of writing (i.e., text length, lexical sophistication, cohesion, and syntactic complexity), which have been found important attributes of quality writing, account for a substantial portion of the variance in holistic judgments of L2 writing proficiency. Similar pictures have been obtained in previous research conducted in different contexts (e.g., TOEFL essays; Guo et al., 2013). Thus, the findings suggest that computational indices are capable of measuring L2 writing proficiency across different (high stakes) language tests, although more research is needed to corroborate this notion.

Follow-up studies are warranted that explore whether computational indices correspond to the features of L2 writing proficiency that they purport to measure in other high stakes writing tests. This is especially pertinent to writing samples that are taken from outside the genre of independent writing, which is assessed in the MELAB. Other genres of interest would include integrated writing, research reports, and narratives. Such future research will permit us to better understand not only the role of the automated indices in predicting L2 writing proficiency but also the reliability and validity of the computational indices for evaluating L2 writing proficiency.

Acknowledgements

This paper reports on research funded through CaMLA's Spaan Research Grant Program, 2014.

The Authors

YeonJoo Jung is an assistant professor in the Department of English Education of Pusan National University in South Korea. Her research interests include second language acquisition (SLA) and task-based language teaching. Within SLA, her primary focus is on the application of experimental techniques from psychology to second language processing and acquisition

Department of English Education
College of Education
Pusan National University
Busan, 46241, Korea
Mobile: +82 51-510-1612
Email: yeonjoo.jung@gmail.com

Scott Crossley is a Professor in the Department of Applied Linguistics and ESL of Georgia State University in the USA. His interests include the application and development of natural language processing tools in educational technology. He has published articles on the use of natural language processing tool to examine lexical acquisition, writing proficiency, reading comprehension, discourse processing, language assessment, and automatic feedback in intelligent tutoring systems.

Department of Applied Linguistics and ESL
College of Arts and Sciences
Georgia State University
Atlanta, GA 30303, USA
Tel: +1 404-413-5200
Email address: scrossley@gsu.edu

Danielle McNamara is a Professor in the Department of Psychology of Arizona State University in the USA. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research ranges a variety of topics including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms.

Department of Psychology
College of Liberal Arts and Sciences
Arizona State University
Tempe, 85287 AZ, USA
Tel: +1 480-965-7598
Email address: dsmcnamara1@gmail.com

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (Eds.). (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium.

Barkaoui, K. (2007). Teaching writing to second language learners: Insights from theory and research. *TESL Reporter, 40*, 35-48.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics, 21*, 543-565.

British National Corpus, version 3 (BNC XML ed.). (2007). Retrieved from http://www.natcorp.ox.ac.uk

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instrumentation & Computers, 16*, 502-532.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904-911.

Buckwalter, J. K., & Lo, Y. G. (2002). Emergent biliteracy in Chinese and English. *Journal of Second Language Writing, 11*, 269-293.

CaMLA. (2013). *The MELAB 2013 Report*. Ann Arbor, MI: Cambridge Michigan Language Assessment.

Chenoweth, N., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication, 18*, 80-98.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater®'s performance on TOEFL essays* (TOEFL Research Report No. 73). Princeton, NJ: Educational Testing Service.

Connor, U. (1990). Linguistic/rhetorical measures of international persuasive student writing. *Research in the Teaching of English, 24*, 67-87.

Crossley, S. A., Kyle, K., Allen, L., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation. *Journal of Writing Assessment, 7*(1). Retrieved from http://www.journalofwritingassessment.org/article.php?article=74

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The Tool for the Automatic Analysis of Text Cohesion (TASCO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48,* 1227-1237. .

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In P. M. McCarthy & G. M. Youngblood (Eds.). *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 208-213). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29,* 243-263.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27,* 317-334.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

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

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

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

Kuperman, V., Stadthagen-Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44*, 978-990.

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

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Lee, J. J., Bychkovska, T., & Maxwell, J. D. (2018). Breaking the rules? A corpus-based comparison of informal features in L1 and L2 undergraduate student writing. *System, 80*, 143-153.

Louwerse, M. M., & Graesser, A. C. (2005). Coherence in discourse. In P. Strazny (Ed.), *Encyclopedia of linguistics*. (pp. 216-218). Chicago, IL: Fitzroy Dearborn.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313-330.

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.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher. 18*, 5-11.

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.

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective* (Unpublished doctoral dissertation). Northern Arizona University-Flagstaff.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly, 27*, 657-677.

Toglia, M. P., & Battig, W. R. (1978). *Handbook of semantic word norms.* New York: Erlbaum.

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*, 85-99.

Whitten, I. A., & Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Elsevier.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instrumentation and Computers, 20*, 6-10.

Appendix

Correlations between Essay Scores and Selected Indices

Index	Category	*r*
Number of words	Text structure	0.618
BNC Written Bigram Freq Log	Lexical sophistication	0.509
Word familiarity (content words)	Lexical sophistication	-0.506
Word frequency (content words_spoken)	Lexical sophistication	-0.493
BNC Written Freq CW Log	Lexical sophistication	-0.485
Lexical Diversity *D* (all words)	Lexical sophistication	0.469
Kuperman AoA CW Log	Lexical sophistication	0.468
Lexical diversity MTLD (all words)	Lexical sophistication	0.450
Word meaningfulness (all words)	Lexical sophistication	-0.449
KF Ncats CW	Lexical sophistication	-0.445
CELEX Word frequency (content words)	Lexical sophistication	-0.438
Incidence of attributive adjective	Syntactic complexity	0.434
Incidence of noun phrase	Syntactic complexity	-0.423
Brown Freq CW Log	Lexical sophistication	-0.416
Adjacent overlap two paragraphs FW lemma average	Cohesion	0.412
Content word overlp (all words_mean)	Cohesion	-0.402
Brysbaert Concreteness FW	Lexical sophistication	-0.395
Incidence of logical operator	Cohesion	-0.386
Lexical diversity, Maas (all words)	Lexical sophistication	-0.377
Lexical diversity, K (all words)	Lexical sophistication	-0.374
Part of speech: Verb non-3rd person singular present	Syntactic complexity	-0.346
Adjacent overlap lemma CW	Cohesion	-0.345
Number of syllables per word	Lexical sophistication	0.340
Verb synonym paragraph lemma overlap	Cohesion	0.335
Argument overlap (all words_mean)	Cohesion	-0.327
Narativity	Text structure	-0.326
BNC Written Trigram Freq Normed (word) Log	Lexical sophistication	0.325
Number of sentences	Text structure	0.324
Number of logical connectives	Cohesion	-0.310
MRC Imageability FW	Lexical sophistication	-0.302
Brysbaert Concreteness Combined AW	Lexical sophistication	-0.294
Adjacent overlap binary two paragraphs noun lemma average	Cohesion	0.292
Noun synonym paragraph overlap	Cohesion	0.282
Number of modifiers per noun phrase	Syntactic complexity	0.277
Incidence of pronoun	Syntactic complexity	-0.277
Incidence of determiners	Syntactic complexity	0.276
Brysbaert Concreteness Unigram FW	Lexical sophistication	-0.273
Number of positive logical connectives	Cohesion	-0.263
Adjacent overlap sentence verb lemma	Cohesion	-0.263
Adjacent overlap sentence noun lemma	Cohesion	-0.262
Number of positive intentional connectives	Cohesion	-0.258
Number of determiners	Cohesion	0.257
LIWC (space words)	Lexical sophistication	0.252
Incidence of causal connectives	Cohesion	-0.250
Number of subject personal pronoun	Syntactic complexity	-0.244
Incidence of adjective predicate	Syntactic complexity	0.242

Situation model (Zsit)	Cohesion	-0.235
Semantic similarity (LSA sentence to sentence)	Cohesion	-0.234
Number of words before the main verb	Syntactic complexity	0.233
Part of speech: preposition	Syntactic complexity	0.228
Semantic similarity (LSA introduction to conclusion)	Cohesion	0.227
Number of demonstrative	Cohesion	0.221
Present tense	Cohesion	0.217
Adjacent overlap binary two paragraphs verb lemma average	Cohesion	0.208
Semantic similarity (LSA sentence all across paragraphs)	Cohesion	-0.206
Number of basic connectives	Cohesion	-0.203
Word hypernymy (nouns & verbs)	Lexical sophistication	0.192
Number of higher level constituents per word	Syntactic complexity	-0.183
Number of all connectives	Cohesion	-0.177
Adjacent overlap binary two sentences pronoun lemma average	Cohesion	-0.177
Part of speech: to (infinitives)	Syntactic complexity	0.175
Number of positive causal connective	Cohesion	0.159
Type-token ratio for content words	Lexical sophistication	0.159
WordNet verb overlap	Cohesion	0.157
Incidence of adverb	Syntactic complexity	0.152
Semantic similarity (LSA paragraph to paragraph)	Cohesion	0.152
Type-token ratio for bigrams	Lexical sophistication	0.148
Incidence of all conjuncts	Cohesion	0.142
Part of speech: Verb	Syntactic complexity	-0.141
Number of causal connectives	Cohesion	-0.140
Number of simple subordinators	Cohesion	0.138
LIWC (verbs)	Lexical sophistication	-0.136
Number of positive connectives	Cohesion	0.125
Minimal Edit Distance	Cohesion	-0.124
Incidence of causal verbs & causal particles	Cohesion	-0.121
Word polysemy	Lexical sophistication	-0.075*
Incidence of adjective	Syntactic complexity	0.065*

* $p < .05$; all others $p < .01$

Note. AW = all words; CW = content words; FW = function words; Freq = frequency; Log = logarithm