Assessing Text Readability Using Cognitively Based Indices

SCOTT A. CROSSLEY
Mississippi State University
Mississippi State, Mississippi, United States

JERRY GREENFIELD
Miyazaki International College
Miyazaki, Japan

DANIELLE S. McNAMARA
University of Memphis
Memphis, Tennessee, United States

Many programs designed to compute the readability of texts are narrowly based on surface-level linguistic features and take too little account of the processes which a reader brings to the text. This study is an exploratory examination of the use of Coh-Metrix, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis. It is suggested that Coh-Metrix provides an improved means of measuring English text readability for second language (L2) readers, not least because three Coh-Metrix variables, one employing lexical coreferentiality, one measuring syntactic sentence similarity, and one measuring word frequency, have correlates in psycholinguistic theory. The current study draws on the validation exercise conducted by Greenfield (1999) with Japanese EFL students, which partially replicated Bormuth's (1971) study with American students. It finds that Coh-Metrix, with its inclusion of the three variables, yields a more accurate prediction of reading difficulty than traditional readability measures. The finding indicates that linguistic variables related to cognitive reading processes contribute significantly to better readability prediction than the surface variables used in traditional formulas. Additionally, because these Coh-Metrix variables better reflect psycholinguistic factors in reading comprehension such as decoding, syntactic parsing, and meaning construction, the formula appears to be more soundly based and avoids criticism on the grounds of construct validity.

Accurately predicting the difficulty of reading texts for second language (L2) learners is important for educators, writers, publishers, and others to ensure that texts match prospective readers' proficiency. This study explores the use of Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, & Graesser, 2002), a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis, as an improved means of measuring English text readability for L2 readers.

Although traditional readability formulas such as Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid, Fishburne, Rogers, & Chissom, 1975) have been accepted by the educational community, they have been widely criticized by both first language (L1) and L2 researchers for their inability to take account of deeper levels of text processing (McNamara, Kintsch, Butler-Songer, & Kintsch, 1996), cohesion (Graesser et al., 2004; McNamara et al., 1996), syntactic complexity, rhetorical organization, and propositional density (Brown, 1998; Carrell, 1987). Coh-Metrix offers the prospect of enhancing traditional readability measures by providing detailed analysis of language and cohesion features through integrating language metrics that have been developed in the field of computational linguistics (Jurafsky & Martin, 2000). Coh-Metrix is also well suited to address many of the criticisms of traditional readability formulas because the language metrics it reports on include text-based processes and cohesion features that are integral to cognitive reading processes such as decoding, syntactic parsing, and meaning construction (Just & Carpenter, 1987; Perfetti, 1985; Rayner & Pollatsek, 1994).

L1 READABILITY

Providing students with texts that are accessible and well matched to reader abilities has always been a challenge for educators. A solution to this problem has been the creation and use of readability formulas. Since 1920, more than 50 readability formulas have been produced in the hopes of providing tools to measure text difficulty more accurately and efficiently. The majority of these formulas are based on factors that represent two broad aspects of comprehension difficulty: (a) lexical or semantic features and (b) sentence or syntactic complexity (Chall & Dale, 1995). According to Chall and Dale, formulas that depend on these variables are popular because they are easily associated with text simplification. For instance, a text written for early readers generally contains more frequent words and shorter sentences. Thus, on an intuitive level, measuring the word frequency and sentence length of a text should provide a basis for understanding how readable the text is.

A number of first language validation studies have found the predictive validity of traditional readability formulas to be high, correlating with observed difficulty in the $r = 0.8$ range and above (Chall & Dale, 1995). Traditional readability formulas, however, are generally not based on theories of reading or comprehension building, but on tracing statistical correlations. Therefore, the credibility accorded to them is strictly based on their demonstrated predictive power, and they are often accused of having weak construct validity. The limited validity of the formulas has inclined many researchers within the field of discourse processing to regard them with reservation and to caution against their use (Davison & Kantor, 1982; Rubin, 1985). However, the attraction of simple, mechanical assessments has led to their common use for assessing all sorts of text designed for a wider variety of readers and reading situations than those for which the formulas were created.

The shortcomings of traditional formulas also become evident when one matches them against psycholinguistic models of the processes that the reader brings to bear on the text. Psycholinguists regard reading as a multicomponent skill operating at a number of different levels of processing: lexical, syntactic, semantic, and discoursal (Just & Carpenter, 1987; Koda, 2005). It is a skill that enables the reader to make links between features of the text and stored representations in his or her mind. These representations are not only linguistic, but include world knowledge, knowledge of text genre, and the discourse model which the reader has built up of the text so far. The reader can also draw on multiple previous reading experiences which have finely tuned the processes that are brought to bear on a text; in the case of the L2 reader, of course, the processes will have been acquired in relation to the L1 and undergone adaptation (perhaps incomplete) to the rather different circumstances of reading in the L2.

Clearly, a psycholinguistically based assessment of text comprehensibility must go deeper than surface readability features to explain how the reader interacts with a text. It must include measures of text cohesion and meaning construction (Gernsbacher, 1997; McNamara et al., 1996) and encode comprehension as a multilevel process (Koda, 2005). This encoding would include, inter alia, measures related to decoding, syntactic parsing, and meaning construction (Just & Carpenter, 1987; Perfetti, 1985; Rayner & Pollatsek, 1994). In due course, a readability measure would need to be framed that takes appropriate account of the role of working memory and the constraints it imposes in terms of propositional density and complexity.

TRADITIONAL READABILITY FORMULAS FOR L2 READERS

A number of studies have examined the relationship between traditional readability formulas (e.g., Flesch-Kincaid grade level, Kincaid et al., 1975; Flesch reading ease, Flesch, 1948) and L2 evaluations of readability and text difficulty. These studies were undertaken because researchers were dissatisfied with classic readability formulas when applied to text design for L2 readers. Like traditional L1 readability formulas, those used in L2 have generally depended on surface-level sentence difficulty indices, such as the number of words per sentence and surface-level word difficulty indices such as syllables per words (Brown, 1998; Greenfield, 1999).

Carrell (1987) discussed both the importance of developing an accurate L2 readability measure and the faults of traditional readability formulas when applied to L2 texts. She argued that more accurate readability formulas were needed to ensure a good match between L2 reading texts and L2 learners. She was critical of traditional readability formulas for not accounting for reader characteristics or for text-based factors such as syntactic complexity, rhetorical organization, and propositional density. Brown (1998) was also concerned that traditional readability formulas failed to account for L2 reader-based variables. In addition, he argued that readability formulas for L2 readers needed to be sensitive to the type, function, and frequency of words and to word redundancy within the text.

Remarkably, in spite of this concern within the field of L2 reading, little attention has been given to the empirical validation of traditional readability formulas in relation to L2 contexts. Even less has been given to developing alternatives more in line with current knowledge about psycholinguistic models of L1 or L2 reading. Most, if not all, studies that have investigated readability formulas for L2 students have depended on traditional readability measures (e.g., Brown, 1998; Greenfield, 1999, 2004; Hamsik, 1984). Hamsik's study, for instance, examined Flesch-Kincaid and other traditional formulas. Hamsik determined that the formulas "do measure readability [for] ESL students and that they can be used to select material appropriate to the reading level of ESL students" (p. iv). However, Hamsik's study was neither large enough nor sufficiently fine grained to settle the question of predictive validity (Greenfield, 1999), nor did it consider cognitive factors.

Brown (1998) examined the validity of traditional readability formulas for L2 learners using 12th-word cloze procedures[1] on passages from 50 randomly chosen English adult reading books read by 2,300 Japanese EFL learners. He compared the observed mean cloze scores on the passages with scores predicted by six readability measures, including the Flesch and Flesch-Kincaid. The resulting correlations ranged from 0.48–0.55, leading him to conclude that "first language readability indices are

[1] Cloze procedures involve the systematic deletion of words in text. Text comprehension is measured by how accurately the reader can insert an acceptable word into the deleted slot. The validity of this method has been widely debated, but not conclusively resolved (Oller & Jonz, 1994). It is, however, a durable difficulty criterion that has been used in multiple L1 and L2 readability studies (Bormuth, 1969; Chall & Dale, 1995; Greenfield, 1999).

not very highly related to the EFL difficulty" (p. 27). Using multiple regression analyses on a training set only, Brown then created a new readability formula by selecting variables that were more highly predictive of difficulty for L2 readers. Brown's EFL readability index comprises a small subset of variables that include the average number of syllables per sentence, the frequency of the cloze items in the text as a whole, the percentage of words in the text of more than seven letters, and the percentage of function words. With a multiple correlation of 0.74 and an R^2 of 0.51, Brown's formula demonstrated a higher degree of association and accounted for more variance in his L2 learners' scores than did the traditional formulas.

Greenfield (1999) analyzed the performance of 200 Japanese university students on the set of academic passages used in Bormuth's (1971) readability study. Following Bormuth's methodology, he constructed fifth-word deletion cloze tests on 31 of Bormuth's 32 passages (one passage was read by all participants as a control, and one was omitted for a balanced design). Pearson correlations between the observed mean cloze scores of the Japanese students and the scores predicted by traditional readability formulas ranged from 0.69 for the New Dale-Chall formula (Dale & Chall, 1995) to 0.85 for Flesch reading ease (Flesch, 1948) and Flesch-Kincaid (Kincaid et al., 1975), and 0.86 for Bormuth (1971).

Greenfield (1999) next used the set of mean cloze scores to examine whether a regression with traditional readability variables would yield a significant improvement over the traditional formulas in predicting the scores of the EFL readers. A comprehensive check of all of the classic variables found that a regression of just two surface-level variables, letters per word and words per sentence, against the observed mean scores produced an EFL difficulty index that was as good as or slightly better than any of the classic formulas, with a multiple correlation of $R = 0.86$, $R^2 = 0.74$, and adjusted $R^2 = 0.72$.[2] The new formula, called the Miyazaki EFL readability index, had the advantage of being scaled for L2 readers.

Comparing his own study with Brown (1998), Greenfield (2004) argued that Brown's passage set was not sufficiently variable in difficulty and too difficult overall to provide a measure of L2 reading. Greenfield also regressed Brown's variables against the Miyazaki EFL reading scores[3]

[2] Note that Greenfield (1999) reported an adjusted R^2 whereas Brown (1998) reported an R^2. An adjusted R^2 is different from an R^2 because it estimates the loss of predictive power. It is a more conservative measure and is used as an estimate of cross-validation.

[3] However, as Greenfield acknowledged, it is likely that the model was overfitted for the Miyazaki criterion by applying too many independent variables (four: syllables per sentence, passage frequency, long words, and function words) against a dependent variable with too few cases (30 passages). An overfitting such as this leads to findings that are statistically questionable because when data samples are regressed against too many variables, random data can appear to show a strong effect.

and found a multiple correlation of $R = 0.91$, $R^2 = 0.83$, and adjusted $R^2 = 0.79$

The studies that have been mentioned offer some evidence that classic readability measures discriminate relative difficulty reasonably well for L2 students. They only appear to do so when operating on appropriate academic texts for which they were designed, but not to the level of accuracy achieved in L1 cross-validation studies. Adjustment of the classic model based on an EFL readability score offers only slight improvement (Greenfield, 1999). The possibility arises that constructing a new model incorporating at least some variables that reflect the cognitive demands of the reading process may yield a new, more universally applicable measure of readability.

COH-METRIX

As reported in Graesser et al. (2004), recent advances in numerous disciplines have made it possible to computationally investigate various measures of text and language comprehension that supersede surface components of language and instead explore deeper, more global attributes of language. The various disciplines and approaches that have made this approach possible include psycholinguistics, computational linguistics, corpus linguistics, information extraction, information retrieval, and discourse processing. Taken together, the advances in these fields have allowed the analysis of many deep-level factors of textual coherence and processing to be automated, permitting more accurate and detailed analyses of language to take place.

A synthesis of the advances in these areas has been achieved in Coh-Metrix, a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis. This tool was designed with the goal of improving reading instruction by providing a means to guide textbook writing and to match textbooks more appropriately to the intended students (Graesser et al., 2004). Coh-Metrix represents an advance on conventional readability measures such as Flesch-Kincaid and Flesch reading ease because it reports on detailed language and cohesion features. The system integrates semantic lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters, and other components that have been developed in the field of computational linguistics (Jurafsky & Martin, 2000). This integration allows for the examination of deeper level linguistic features of text that are related to text processing and reading comprehension.

The purpose of this study was to examine if certain Coh-Metrix variables can improve the prediction of text readability. Implicit within this purpose was the examination of variables that more accurately reflect the cognitive processes which contribute to skilled L2 reading. It was hypothesised that an analysis of variables relating to lexical frequency, syntactic similarity, and content word overlap would allow for an improved measure of readability. The significance of these variables is that they broadly correspond. respectively, to three operations which many psycholinguistic models of reading and comprehension distinguish: decoding, syntactic parsing, and meaning construction.

METHOD

Materials

Bormuth's (1971) corpus of 32 academic reading texts was selected to test the hypothesis that linguistic variables related to cognitive processing and cohesion could better predict text readability. The Bormuth reading set features texts taken from school instructional material and includes passages from biology, chemistry, civics, current affairs, economics, geography, history, literature, mathematics, and physics. The mean length of the texts was 269.28 words (*SD* = 16.27), and the mean number of sentences per hundred words was 7.10 (*SD* = 2.81). The process of selection was informed by the work of Chall and Dale (1995), who evaluated the Bormuth passages for text characteristics and cross-validation of readability scores and found them more advantageous than other available passage sets. More important, as discussed earlier, Greenfield (1999) used 31 of the Bormuth passages to test the reading skills of Japanese university-level Japanese students, collecting scores based on fifth-word deletion cloze tests.

In this study, we used the same passage set and the same mean cloze scores taken from the 200 Japanese participants studied by Greenfield (1999). We also conducted similar statistical analyses to Greenfield except that we measured readability using cognitively based variables related to reading processes. While we recognize the limitations found in the size of the passage set and in the scoring criterion, we also recognize that the passage set has served as a basis for two classic studies (Bormuth, 1971; Chall & Dale, 1995) and a recent cross-validation with an EFL population sample (Greenfield).

Variable Selection

Independent variables to measure text readability were chosen from existing Coh-Metrix banks of indices based on a-priori assumptions taken from the L1 and L2 reading literature. The number of passages available

(31 in this case) limited the number of predictors that could safely be used without overfitting the model. Generally, a minimum of 10 cases of data for each predictor is considered to be accurate (with conservative models using 15 to 20). Accordingly, three banks of indices were selected to analyze the Bormuth passages. The indices that were selected correspond to three general levels into which many psycholinguistic accounts divide reading, namely, lexical recognition, syntactic parsing, and meaning construction (Just & Carpenter, 1987; Perfetti, 1985; Rayner & Pollatsek, 1994).

Lexical Index

Coh-Metrix calculates word frequency information through CELEX frequency scores. The CELEX database (Baayen, Piepenbrock, & Gulikers, 1993) consists of frequencies taken from the early 1991 version of the COBUILD corpus, a 17.9 million-word corpus. For this study, the *CELEX frequency score for written words* was selected as the lexical-level variable. This measure was selected because frequency effects have been shown to facilitate decoding. Frequent words are processed more quickly and understood better than infrequent ones (Haberlandt & Graesser, 1985; Just & Carpenter, 1980). Rapid or automatic decoding is a strong predictor of L2 reading performance (Koda, 2005). Texts which assist such decoding (e.g., by containing a greater proportion of high-frequency words) can thus be regarded as easier to process.

Syntactic Index

The index *semantic similarity: sentence to sentence, adjacent, mean* measures the uniformity and consistency of parallel syntactic constructions in text. The index not only looks at syntactic similarity at the phrase level, but also takes account of the parts of speech involved, on the assumption that the more uniform the syntactic constructions are, the easier the syntax will be to process. It is important to include a measure of difficulty that is not simply based on the traditional L2 grading of grammar patterns but also takes account of how the reader handles words as they are encountered on the page. A reading text is processed linearly, with the reader decoding it word by word; but, as he or she reads, the reader also has to assemble decoded items into a larger scale syntactic structure (Just & Carpenter, 1987; Rayner & Pollatsek, 1994). Clearly, the cognitive demands imposed by this operation vary considerably according to how complex the structure is (Perfetti, Landi, & Oakhill, 2005). They also vary according to how predictable the final part of the structure is because, while still in the course of reading a sentence, we form expectations as to how it will end. So-called garden path sentences such as *John* *remembered the answer / was in the book.* impose particularly heavy demands and contribute significantly to text difficulty (Field, 2004, pp. 121, 299). These factors of potential difficulty are provided for by the Coh-Metrix semantic similarity index

Meaning Construction Index

The Coh-Metrix index *content word overlap,* which measures how often content words overlap between two adjacent sentences, measures one of many factors that facilitate meaning construction. It was selected because overlapping vocabulary has been found to be an important aspect in reading processing and can lead to gains in text comprehension and reading speed (Douglas, 1981; Kintsch & van Dijk, 1978; Rashotte & Torgesen, 1985).

STATISTICAL ANALYSIS

To calculate the readability of the Bormuth passage set, the three selected variables were used as predictors in a training set. A multiple regression equation with the 31 observed EFL scores as the dependent variable was conducted.[4] A limited data set presents a challenge of how to make the most of the available data. Past studies have reported the adjusted R^2 (Greenfield, 1999, 2004), which estimates variance based on the population from which the data were sampled. However, the adjusted R^2 does not estimate how well the model would predict scores of a different sample from the population. To address this problem, this study also reports Stein's unbiased risk estimate (SURE). However, neither approach can estimate how well the model would perform on a separate test set. For this estimate, a technique known as repeated cross-validation is needed. In cross-validation, a fixed number of folds, or partitions of the data, is selected. Once the number of folds has been selected, each is used for testing and training in turn. In *n*-fold cross-validation, one instance in turn is left out and the remaining instances are used as the training set (in this case 30). The accuracy of the model

[4] In perfect circumstances, a researcher would have enough data available to create separate training and testing sets and use the training set to create predictors and the testing set to calculate how well those predictors function independently. Historically, most readability studies have been statistically flawed in that they have based their findings on the results of a single training set. Although performance on a single training set allows conclusions regarding how well variables predict the difficulty of the texts in that set, those conclusions may not be extendible to an independent test set (Whitten & Frank, 2005). The problem, of course, is the difficulty of creating sufficiently large data sets. With only 50 passages in Brown's (1998) study and 30 passages in Greenfield's (1999), it was not feasible for either of them to create both training and test sets.

is tested on the model's ability to predict the omitted instance. In the case of the data at hand, predictors were taken from the training set and used in a regression analysis of the first 30 texts. The *B* values and the constant from that analysis were used to predict the value of the 31st text. This process was repeated for all 31 texts, creating a testing set. The predicted values were then correlated with the actual values (the mean cloze scores) to test the model for performance on an independent testing set.

All of these models (adjusted R^2, SURE estimate, and *n*-fold cross-validation) are important, because if a model can be generalized, then it is likely capable of accurately predicting the same outcome variable from the same set of predictors in a different text group. Thus, if the models are significant, by extension, we can argue that the readability formula would be successful in textual genres other than academic texts.

RESULTS

Correlation and Multiple Regression

In order to estimate the degree to which the chosen independent variables were collectively related to predicting the difficulty of the Bormuth passages for EFL readers, the dependent and independent variables were investigated using multiple regression. A stepwise multiple regression analysis was calculated for the three variables regressed against the mean EFL cloze scores for the Bormuth passages. Descriptive statistics for the dependent and independent variables appear in Table 1, and results for the regression analysis appear in Table 2.

The multiple regression analysis also reported individual Pearson correlations for each selected variable. When comparing the three selected variables to the EFL mean cloze scores, significant correlations were obtained for all indices. Correlations between the Bormuth mean cloze scores and the adjacent sentence similarity score were significant ($n = 31$, $r = 0.71$, $p < 0.001$), as was the content word overlap score ($n = 31$, $r =$

TABLE 1
Descriptive Statistics

Variable	Mean	Standard deviation	N
Predicted			
Mean cloze scores	23.854	12.944	31
Predictor			
Content word overlap	0.1457	0.090	31
Sentence syntax similarity	0.149	0.087	31
CELEX frequency	2.349	0.243	31

TABLE 2
Stepwise Regression Analysis of Three Independent Variables Predicting EFL Reading Difficulty

Dependent variable: EFL difficulty

Step 1	$R = 0.793$	$R^2 = 0.628$	Added Content word overlap
Step 2	$R = 0.887$	$R^2 = 0.786$	Added Sentence syntax similarity
Step 3	$R = 0.925$	$R^2 = 0.856$	Added CELEX frequency

Variable	Unstandardized coefficient	Standardized coefficient	Standard error	T	p
Content word overlap	52.230	0.362	16.827	3.104	0.004
Sentence syntax similarity	61.306	0.412	17.030	3.600	0.000
CELEX frequency	22.205	0.416	4.054	5.477	0.000

Note. Estimated Constant Term = –45.032

0.79, $p < 0.001$), and the CELEX written frequency score ($n = 31$, $r = 0.61$, $p < 0.001$). These correlations demonstrate the strength of the linear relationships between the mean readability scores and the individual variables.

The results of the multiple regression analysis indicate that the combination of content word overlap, syntactic similarity, and CELEX frequency taken together produce a multiple correlation of 0.93 and a corresponding R^2 of 0.86. Translated, this result signifies that the combination of the three variables alone accounts for 86% of the variance in the performance of the Japanese students on the 31 cloze tests based on the Bormuth passages. In other words, using these three variables, the model can predict 86% of the difficulty for these passages.

Cross-Validation

Three estimates of cross-validation were conducted as described in the previous section. The adjusted R^2 for the regression analysis is 0.84 and the SURE is 0.81. Inasmuch as these two estimates are very similar to the observed R^2 (0.86), the cross-validity of the model is supported. In the n-fold cross-validation model, the correlation between the predicted values of the testing set and the actual values reveals a significant correlation ($n = 31$, $r = 0.91$, $p < 0.001$), demonstrating that the predictors would likely perform well on an independent testing set.

Comparison With Other Measures

Predictions made by the new index were compared with those made by the Flesch reading ease, Flesch-Kincaid grade level (Bormuth, 1969),

TABLE 3
Pearson Correlations Between Observed Scores and Scores Predicted by Various Readability Measures

Readability measure	Observed EFL
Flesch reading ease	–0.845
Flesch-Kincaid grade level	–0.847
Bormuth formula	0.861
Dale-Chall formula	0.691
Miyazaki EFL index	0.848
Coh-Metrix EFL index	0.925

Note. All relationships significant at $p < 0.01$, $n = 31$. The negative sign can be ignored as an artifact of contrasting scales.

and Miyazaki formulas. The results are shown in Table 3. In these correlations it is readily apparent that the predictions made by the Coh-Metrix EFL index are stronger than those made by any of the other formulas. To confirm that the differences are statistically significant, a Williams *t* test of related correlations was applied. The results of this test, shown in Table 4, indicate that the Coh-Metrix formula has a clear superiority in accuracy to all of the other indices.

DISCUSSION

This study has investigated whether a new tool for measuring text cohesion and text difficulty (Coh-Metrix) might provide a measure of text difficulty that is more accurate than traditional readability formulas are for readers of English as a second or foreign language. Predicting accurate readability of text is crucial to ensure that the input to which L2 readers are exposed matches their processing ability and allows for the noticing, comprehension, and intake of the L2. Using Greenfield's

TABLE 4
Williams *t* Test: Coh-Metrix Formula Versus Classic and Miyazaki EFL Formulas

Formula	Coh-Metrix EFL vs. observed EFL	Formula vs. observed EFL	Coh-Metrix EFL vs. formula	Williams *t* value*
Flesch reading ease	0.925	–0.845	0.898	7.642
Flesch-Kincaid grade level	0.925	–0.847	0.918	10.146
Bormuth formula	0.925	0.861	0.939	12.549
Dale-Chall formula	0.925	0.691	0.704	6.063
Miyazaki EFL index	0.925	0.848	0.969	41.299

Note. Underlined values are significantly larger in that row's comparison. All values significant at $p < 0.01$; $t = 2.762$ needed for significance at $p < 0.01$ (2-tailed, $df = 28$). The negative sign can be ignored as an artifact of contrasting scales.

(1999) L2 mean cloze reading scores, it was found that three Coh-Metrix variables broadly related to the cognitive processes that readers use yielded a more accurate prediction of reading difficulty of Bormuth's (1971) classic passage set than did either Bormuth's own formula or other traditional readability measures. The resulting formula, which may provisionally be called the Coh-Metrix L2 reading index, is as follows:

$$\begin{aligned}\text{Predicted cloze} = -45.032 &+ (52.230 \times \text{Content Word Overlap Value})\\ &+ (61.306 \times \text{Sentence Syntax Similarity Value})\\ &+ (22.205 \times \text{CELEX Frequency Value})\end{aligned}$$

The accuracy of this formula in predicting the EFL difficulty of the Bormuth passages for Japanese readers is statistically superior to that of any of the traditional formulas. This result is encouraging not only because of the strength of the multiple correlation found, but also because the improvement seems in part to be attributable to the incorporation of variables related to the conventional psycholinguistic operations of decoding, syntactic parsing, and meaning construction.

The readability formula presented here is exploratory and only considers three indices out of the hundred or so available through Coh-Metrix. These additional indices will allow future researchers options for incorporating measurements of cohesion such as anaphoric resolution, temporal and spatial information, semantic information, and indices of causality, to name but a few. All of these features have a cognitive dimension as well as a linguistic one, and they are important in psycholinguistic accounts of reading.

APPLICATIONS

Most important, however, this study has immediate transfer potential in that it provides a readability formula that is based on freely accessible computational indices.[5] The use of a cognitively based readability formula that is better suited to predict the readability of L2 texts could provide materials developers and classroom teachers with a valuable resource for analyzing and selecting appropriate text for L2 learners. Past studies using Coh-Metrix (Crossley, Louwerse, McCarthy & McNamara, 2007; Crossley & McNamara, 2008) have demonstrated that L2 reading texts need to be evaluated using more global indices of lan-

[5] To access the variables used in this study visit http://cohmetrix.memphis.edu/cohmetrixpr/index.html. The Web site provides users with values for various linguistic variables related to cognitive processing as well as surface-level variables. In reference to the readability formula reported in this article, the Web site reports values for CELEX frequency values, sentence similarity, and content word overlap.

guage, discourse, and conceptual features to ensure closer matches between reader and text. The readability formula described in this article advances that notion. Materials developers, especially those working with graded and abridged texts, could use the Coh-Metrix L2 reading index to better assess text readability. The indices analyzed in this study could also prove beneficial for materials writers when constructing simplified texts or when adapting authentic texts for an L2 reader. Though the current study only examined the effects of three cognitively linked indices, it provides a strong indication that, when choosing a text, we should not simply rely on surface features such as sentence length or syntactic complexity but should also take account of the processes which a reader brings to the text. Those processes may be affected by features of the text which are not normally considered in readability judgments.

One of the criticisms of classic readability formulas has been directed at their inappropriate use by writers and materials adapters as prescriptive guides rather than measurement tools. This mistake grows out of a failure to understand the difference between formulas as predictors of difficulty and formulas as explanations of difficulty. As explanations, the classic formulas fail. The fact that a formula estimates readability from word and sentence length, for instance, could lead a writer to think that simply using short words in short sentences would make a text easier to read. This, however, is a false premise because word length does not always correlate to word frequency and short sentences often omit cohesive markers and are not always easier to parse.

Although there is nothing new in suggesting that lexical frequency is a criterion that teachers and materials designers need to apply when selecting a text for a particular level of learners, using an automated measure of frequency provides a new approach. Historically, the rationale behind using frequency as a criterion has been that reading texts should reflect the order in which learners are likely to have acquired their vocabulary (vocabulary in reading reflects vocabulary in core course books) and the relative usefulness of the words in question (the more frequent the word, the more likely the learner is to need it). In the discussion of this study a third, cognitive, argument has been added. The more frequent a word, the more likely it is to be processed with a fair degree of automaticity, thus increasing reading speed (even among lower level learners) and freeing working memory for higher level meaning building.

Syntactically, basic graded readers tend to adhere quite closely to a canonical subject–verb–object pattern and to use simple rather than complex syntactic structures, with coordination but minimal subordination. Again, this approach aims to represent what the learner is likely to know. As progress in the L2 is made, the variety of sentence types becomes greater. The measure used in this study draws attention to the impact on reading difficulty not of individual structures but of syntactic variety. It has been presented in terms of the parsing operation that the reader needs to carry out. Teachers and materials writers would thus do well to take due account of syntactic variety as well as the more obvious criterion of syntactic complexity. Selecting syntactic structures that parallel each other would provide important links between sentences. So, instead of concentrating on how elementary or how advanced a syntactic structure appears, material developers might concentrate on how frequently the structure is repeated within the text and how this repetition supports linear processing. Reiterated syntactic structures lower the cognitive demands placed on L2 learners and afford them the opportunity to concentrate on meaning construction.

At the meaning construction level, materials designers might also consider word overlap between text segments. Despite work done by Hoey (1991), readability formulas tend not to account for the strength of repeated lexical items to bind text together. Although we realize that this is only one of many possible indices connected to the building of a discourse representation, this feature is important because it measures lexical links between different elements of the text and can assist in the construction of larger patterns of meaning. The effects of limiting the range of vocabulary in a text are well attested in the literature on graded reading; we argue in this article for a shift in perspective from considering the text to considering the reader. If we measure the extent to which particular lexical items are repeated throughout a particular paragraph rather than simply examining the variety of lexical features present, then we can partly gauge the cognitive demands that the text places on the reader in terms of welding the sentences of the paragraph into an overall representation.

For the classroom teacher, the Coh-Metrix L2 reading index could be used to help select texts for classroom activities taken from real-world sources. It would allow L2 teachers to identify authentic texts that contain appropriate linguistic variables matched to reader level in a way that assists in both ease of comprehension and conceptual processing. This readability formula would also prove valuable to textbook publishers wishing to populate an L2 reading textbook with authentic text passages. An automated readability formula that reports on frequency data, syntactic similarity, and word overlap would prove beneficial in helping to select authentic texts that matched reader proficiency level.

LIMITATIONS OF THE STUDY

As in all studies, this one has limitations. First, there is the question of the testing method. Although cloze procedures have been an enduring feature of most readability studies, their validity has been debated (Oller & Jonz, 1994). Unlike other methods that examine global text understanding such as recall and comprehension tests, cloze scoring generally assesses readability at the sentence and word level. Thus, cloze scores might even correlate highly to traditional readability formulas (Crossley, Dufty, McCarthy, & McNamara, 2007). Future studies would benefit from triangulation that draws on two or three criteria measures of text comprehensions.

Second, there is the question of the passage set, which were all academic texts taken from secondary textbooks. Any formula based on such a set can be expected to work with similar texts (i.e., academic), but may or not work with other types of text. This study also examined a relatively small passage set, albeit one that has served as a basis for two classic studies (Bormuth, 1971; Chall & Dale, 1995) and a recent cross-validation with an EFL population sample (Greenfield, 1999). However, future studies would do well to use a larger passage set, together with a separate and comparable training set.

Third, in Greenfield's (1999) administration, each passage was read by only 20 readers. Although group scores were statistically checked, the small sample size could have weakened the finding of parity between groups, introducing the possibility of error in the relative difficulty found for one or more passages. It will be desirable to design future studies for a larger population sample in order to avoid such issues.

CONCLUSION

From a practical perspective, the findings of this study provide L2 materials designers, textbook publishers, and teachers with the opportunity to match text better to reader. These findings could allow texts to be better tailored to students' needs and could ensure that the texts use language in a way that takes due account of the way in which the reader is likely to process it. Ultimately, the theoretical goal of English readability research is to devise a measure that has strong construct validity as well as predictive validity and that is sensitive both to text type and to the reader's L1 background. Although traditional readability formulas have generally been found to perform well with academic texts, and recent research suggests that they work adequately in L2 settings, this present study shows that new variables measured by Coh-Metrix contribute significantly to formula accuracy. Because the Coh-Metrix variables selected for this study better reflect certain cognitive operations underlying reading (viz., decoding, syntactic parsing, and meaning construction) than do the surface variables used in traditional formulas, the Coh-Metrix L2 reading index goes some way toward countering accusations of poor construct validity. For now, it remains to be shown whether the Coh-Metrix model will perform as well with other types of text, environments, and occasions for reading and reader populations. There is reason to expect that it will.

ACKNOWLEDGMENTS

The research was supported in part by the Institute for Education Sciences (IES R3056020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors also express their gratitude and debt to Dr. Philip McCarthy, who assisted with the methodologies used in this article, and to the anonymous reviewers and Dr. John Field, who greatly assisted in the final version of this article.

THE AUTHORS

Scott Crossley is an assistant professor and director of the TESOL program at Mississippi State University, United States. His interests include computational linguistics, corpus linguistics, discourse processing, and discourse analysis. He has published articles in genre analysis, cognitive science, multidimensional analysis, speech act classification, and text linguistics.

Jerry Greenfield is a professor at Miyazaki International College in Miyazaki, Japan, where he teaches English as a foreign language, North American art history, and environmental aesthetics. His research interests include L2 text readability, machine–human interfaces, mental models, and bioethics.

Danielle McNamara is a cognitive scientist and director of the cognitive program at the University of Memphis, Memphis, Tennessee, United States. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research focuses on text comprehension and the effects of text coherence.

REFERENCES

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

Bormuth, J. R. (1969). *Development of readability analyses* (Final Report, Project No. 7–0052, Contract No. 1, OEC-3–7-070052–0326). Washington, DC: U. S. Office of Education.

Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance.* U. S. Department of Health, Education and Welfare (ERIC Doc. No. ED O54 233).

Brown, J. D. (1998). An EFL readability index. *JALT Journal, 29*(2), 7–36.

Carrell, P. (1987). Readability in ESL. *Reading in a Foreign Language, 4,* 21–40.

Chall, J., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula.* Cambridge, MA: Brookline Books.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In D. S. McNamara & G. Trafton (Eds.*), Proceedings of the 29th annual conference of the Cognitive Science Society* (pp. 197–202). Austin, TX: Cognitive Science Society.

Crossley, S. A., Louwerse, M. L., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal, 91*(2), 15–30.

Crossley, S. A., & McNamara, D. S. (2008). Assessing second language reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy, & McNamara (2007). *Language Teaching, 41,* 409–429.

Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17,* 187–209.

Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.), *Learning to read in different languages* (pp. 33–102). Washington, DC: Center for Applied Linguistics.

Field, J. (2004). *Psycholinguistics: The key concepts.* New York, Routledge.

Flesch, R. (1948). A new readability yardstick. *The Journal of Applied Psychology, 32,* 221–233.

Gernsbacher, M. (1997). Coherence cues mapping during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships. Studies in the production and comprehension of text* (pp. 3–22). Mahwah NJ: Erlbaum.

Graesser, A. C., McNamara, D. D., Louwerse, M. L., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36,* 193–202.

Greenfield, G. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Unpublished doctoral dissertation, Temple University, Philadelphia, PA, United States. (University Microfilms No. 99–38670).

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal, 26,* 5–24.

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology. General, 114,* 357–374.

Hamsik, M. J. (1984). *Reading, readability, and the ESL reader.* Unpublished doctoral dissertation, University of South Florida.

Hoey, M. (1998). *Pattern of lexis in text.* Oxford: Oxford University Press.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Upper Saddle River, NJ: Prentice-Hall.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329–354.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension.* Boston: Allyn & Bacon, Inc.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel,* Research Branch Report 8–75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363–394.

Koda, K. (2005). *Insights into second language reading.* Cambridge: Cambridge University Press.

McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43.

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). Coh-Metrix (Version 2.0) [Software]. Memphis, TN: University of Memphis, Institute for Intelligent Systems. Available from http://cohmetrix.memphis.edu/cohmetrixpr/index.html

Oller, J. W., & Jonz, J. (1994). *Cloze and coherence.* Cranbury, NJ: Bucknell University Press.

Perfetti, C. A. (1985). *Reading ability.* Oxford: Oxford University Press.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford: Blackwell.

Rashotte, C. A., & Torgesen, J. K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly, 20,* 180–188.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading.* Englewood Cliffs, NJ: Prentice Hall.

Rubin, A. (1985). How useful are readability formulas? In J. Osborn, P. T. Wilson, & R. C. Anderson (Eds.), *Reading education: Foundations for a literate America* (pp. 61–77). Lexington, MA: Lexington Books.

Whitten, I. A., & Frank, E. (2005). *Data mining.* San Francisco: Elsevier.