Reading in a Foreign Language
ISSN 1539-0578

April 2011, Volume 23, No. 1
pp. 84–101

Text readability and intuitive simplification: A comparison of readability formulas

Scott A. Crossley
Georgia State University
United States

David B. Allen
University of Tokyo
Japan

Danielle S. McNamara
University of Memphis
United States

Abstract

Texts are routinely simplified for language learners with authors relying on a variety of approaches and materials to assist them in making the texts more comprehensible. Readability measures are one such tool that authors can use when evaluating text comprehensibility. This study compares the Coh-Metrix Second Language (L2) Reading Index, a readability formula based on psycholinguistic and cognitive models of reading, to traditional readability formulas on a large corpus of texts intuitively simplified for language learners. The goal of this study is to determine which formula best classifies text level (advanced, intermediate, beginner) with the prediction that text classification relates to the formulas' capacity to measure text comprehensibility. The results demonstrate that the Coh-Metrix L2 Reading Index performs significantly better than traditional readability formulas, suggesting that the variables used in this index are more closely aligned to the intuitive text processing employed by authors when simplifying texts.

Keywords: readability, Coh-Metrix L2 Reading Index, simplification, psycholinguistics, cognitive science, computational linguistics, corpus linguistics

When materials developers want to simplify texts to provide more comprehensible input to second language (L2) learners, they generally have two approaches: a structural or an intuitive approach (Allen, 2009). A structural approach depends on the use of structure and word lists that are predefined by level, as typically found in graded readers. Another approach subsumed under the structural approach is one that uses traditional readability formulas such as Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975) or Flesch Reading Ease (Flesch,

1948). These readability formulas provide an indication of text readability that is based on the word and sentence lengths found in the text. In contrast to a structural approach, an intuitive approach is, by nature, more subjective and depends solely on the author's natural sense of text comprehensibility and discourse processing. Both approaches are common in the development of reading materials for L2 learners (Bamford, 1984; Carrell, 1987; Simensen, 1987; Young, 1999).

Our interest in this study is to examine readability formulas' potential for evaluating a corpus of intuitively simplified news texts (Allen, 2009). We are specifically interested in analyzing differences between traditional readability formulas and readability formulas based on psycholinguistic and cognitive accounts of text processing (i.e., the Coh-Metrix L2 Reading Index; Crossley, Greenfield, & McNamara, 2008). Specifically, our purpose is to examine the potential for readability formulas to distinguish among levels of simplified texts (i.e., beginning, intermediate, and advanced) that have been modified using intuitive approaches in order to evaluate the readability formulas' construct validity and to better understand intuitive text simplification. We predict that traditional readability formulas will be less accurate at classifying intuitive levels of simplification than a readability formula founded on cognitively inspired variables. Our foundation for such an argument is that indices contained within the Coh-Metrix L2 Reading Index will better reflect the intuitive text simplification processes used by materials designers because such processes take into account comprehension factors, meaning construction, decoding, and syntactic parsing. Such a finding would provide support for the use of cognitively inspired readability formulas over traditional readability formulas when simplifying text.

Simplified Texts

The purpose of text simplification is straightforward: to provide the L2 reader with text that is more accessible and more comprehensible. Generally, simplified L2 reading texts are either adapted from authentic texts or written explicitly for the L2 reader. At the linguistic level, simplified texts are largely modified to control the complexity of the lexicon and the syntax (Crossley, Louwerse, McCarthy, & McNamara, 2007; Crossley & McNamara, 2008; Simensen, 1987; Young, 1999). For instance, Crossley and his colleagues found that simplified texts, as compared to authentic texts, contained less sophisticated words (e.g., lower frequency words), less syntactic complexity (e.g., number of constituents per sentence), and greater cohesion (e.g., lexical and semantic co-reference). Publishers and material writers also consider the subject matter of the text, the cultural and background knowledge needed to understand the text, the learner, and the literary merit of the text (Lotherington-Wolosyzn, 1993). However, given that our interest lies in the linguistic features related to text processing and how those features inform readability formulas, this study solely examines linguistic modifications.

Supporters of text simplification maintain that the linguistic modifications common in simplification increase the text's comprehensibility and hence the reader's ability to understand and interact with a text (Goodman & Freeman, 1993). Such positions are bolstered by arguments that simplified texts provide more comprehensible input because they contain less lexical sophistication and increased cohesion through redundancy (Allen & Widdowson, 1979; Crossley & McNamara, 2008; Kuo, 1993). Empirical studies examining the comprehensibility of L2

reading texts also support text simplification. For instance, Long and Ross (1993) found that L2 students who read linguistically simplified text scored significantly higher on multiple-choice items intended to assess comprehension than did those that read the authentic version. This finding was supported in a follow up study (Yano, Long, & Ross, 1994), which similarly demonstrated that simplified texts enhanced text comprehension in comparison to authentic texts. A more recent study conducted by Tweissi (1998) also found that simplification positively affected L2 students' reading comprehension. Overall, empirical studies that analyze the benefits of text types support the notion that the use of simplified input results in more comprehensible language and improved comprehension.

Approaches to Text Simplification

As discussed briefly above, material writers have a few choices when simplifying texts. To evaluate the comprehensibility of a simplified text, writers can depend on word or structure lists, traditional readability formulas, on their natural intuition, or a combination of all three. For this study, we are most interested in intuitive approaches and the use of readability formulas. We discuss these in more depth below.

Intuitive approaches

Intuitive approaches are common in L2 text simplification. Author intuition is influenced by personal beliefs and simple hunches about what makes a text more readable (Lotherington-Wolosyzn, 1993). Under an intuitive approach, the author's experiences as a language teacher, language learner, materials writer, or any combination of these guide the process of simplification and allow the authors to rely on their own subjective approximations of what learners at a particular level should be able to understand (Allen, 2009). Even with recourse to word and structure lists for reference, most authors following a structural approach still report relying mainly on their intuition (Young, 1999). Research by Simensen (1987) also reported that most writers depend on their intuition even when publishers offered advice on how to adapt texts. While it is not known how common intuitive approaches toward simplification are, as compared to structural approaches, Simensen's research along with that of Young (1999), Blau (1980), and Carrell (1987) provide evidence that an intuitive approach might not only be extremely common, but perhaps the most common strategy in L2 text simplification.

The process of intuitive text simplification results in reading texts that are theoretically more comprehensible for beginning level learners. Such comprehensibility is the result of less lexical diversity, less sophisticated words (e.g., words that are less frequent, more familiar, more imaginable, and more concrete), less syntactic complexity, and greater cohesion (e.g., more given information, greater semantic co-referentiality, more noun overlap, and greater causal cohesion) in intuitively simplified texts at the beginning level as compared to intuitively simplified texts at the advanced level. However, intuitive text simplification is not without faults because it can also lead to the creation of texts with greater word ambiguity (e.g., more polysemous words) and less word specificity (e.g., verbs with lower hypernymy scores; refer to Crossley, Allen, & McNamara, in press). As reported by Crossley et al. (in press), these textual features are characteristic of intuitive simplification and can be used to classify intuitively

simplified texts at a level well above chance. The majority of these textual features link to factors that make a text more readable and comprehensible. However, the effects of such text modifications on the readability and comprehensibility of text are far from understood.

Traditional Readability Formulas

Another approach to text simplification is the use of traditional readability formulas (Bamford, 1984; Brown, 1998; Carrell, 1987; Greenfield, 2004). Traditional readability formulas are simple algorithms that measure text readability based on sentence length and word length. They have found success in predicting first language (L1) text readability, but have been widely criticized by discourse analysts (Davison & Kantor, 1982) as being weak indicators of comprehensibility and for not closely aligning with the cognitive processes involved in text comprehension (Crossley, Dufty, McCarthy, & McNamara, 2007; McNamara & Magliano, 2009). Traditional readability formulas have also been faulted in the production of L2 texts because they do not account for reader characteristics or text-based factors such as syntactic complexity, rhetorical organization, and propositional density (Carrell, 1987). Carrell argued that more accurate readability formulas were needed to ensure a good match between L2 reading texts and L2 learners. However, the attraction of simple, mechanical assessments has led to traditional readability formulas' common use for assessing a wide variety of texts, readers, and reading situations beyond those for which the formulas were created (Greenfield, 1999).

A few researchers have examined the potential for traditional readability formulas to explain L2 text difficulty, with contradictory findings. Brown (1998), for instance, examined the validity of traditional readability formulas for L2 learners using cloze procedures on passages from 50 randomly chosen English adult reading books read by 2,300 Japanese learners of English as a foreign language (EFL). Brown compared the observed mean cloze scores for the passages with scores predicted by readability measures including Flesch Reading Ease and Flesch-Kincaid Grade Level. The resulting correlations ranged from .48 to .55, leading Brown to conclude that traditional readability formulas were not highly predictive of L2 reading difficulty.

Later, Greenfield (1999) analyzed the performance of 200 Japanese university students using cloze procedures on a set of 32 academic passages used in Bormuth's (1971) study. Pearson correlations between the observed mean cloze scores of the Japanese students and the scores predicted by traditional readability formulas were .85 for both Flesch Reading Ease and Flesch-Kincaid Grade Level. Greenfield, unlike Brown (1998), thus found that traditional readability formulas were predictive of reading difficulty. Noting the difference between Greenfield's (1999) study and Brown's (1998) study, Greenfield (2004) argued that Brown's (1998) passage set was not sufficiently variable in difficulty and too difficult overall to provide a robust passage set for L2 learners. Overall, these studies offer some evidence that classic readability measures discriminate reading difficulty reasonably well for L2 students, but are limited to the appropriate academic texts for which they were designed and do not reach the level of accuracy achieved in L1 cross-validation studies (Greenfield, 1999).

Psycholinguistic and cognitive models of reading also underscore the limitations of traditional formulas (McNamara & Magliano, 2009; Perfetti, 1985; Rayner & Pollatsek, 1994). These models are premised on the notion that reading comprehension is a multi-component skill

focusing on information processing involving both psycholinguistic and cognitive representations (Just & Carpenter, 1980; Koda, 2005; McNamara & Magliano, 2009). The theories underlying the models necessitate a readability measure that takes account of comprehension factors such as coherence (Gernsbacher, 1997; McNamara, Kintsch, Butler-Songer, & Kintsch, 1996) and meaning construction and cognitive processes such as lexical decoding and syntactic parsing (McNamara & Magliano, 2009; Perfetti, 1985; Perfetti, Landi, & Oakhill, 2005; Rayner & Pollatsek, 1994). Cognitive processes are obliquely accounted for in traditional readability formulas (i.e., word length and sentence length are proxy measures of decoding and syntactic parsing), but they are not directly addressed. Comprehension factors are not accounted for in traditional readability formulas.

Coh-Metrix L2 Reading Index

Recent progress in disciplines such as computational linguistics, corpus linguistics, information extraction, and information retrieval have allowed for the development of readability formulas that include indices that more directly correspond to psycholinguistic and cognitive models of reading (e.g., Crossley, Dufty et al., 2007; Crossley et al., 2008). Progress in these fields affords the computational investigation of text using language variables related to text comprehension, cognitive processes, and other factors that go beyond the surface level features of language measured in traditional readability formulas. A synthesis of these developments can be found in Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) a computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis.

Using Coh-Metrix, Crossley et al. (2008) developed an L2 readability formula that incorporated variables that better reflected the psycholinguistic and cognitive processes of reading. Crossley et al. selected three variables to examine the original reading data used in Greenfield's (1999) study. The variables selected by Crossley et al. included a word overlap index (related to text cohesion and meaning construction), a word frequency index (related to decoding), and an index of syntactic similarity (related to parsing). The word frequency and syntactic similarity indices are more closely associated with important cognitive processing constructs than the indices found in traditional readability formulas (i.e., word length and sentence length). The word overlap index provided a variable that was closely aligned to text comprehension processes (i.e., coherence and meaning construction) and one that had no correlate in traditional readability formulas. A regression analysis using these three variables and Greenfield's reading criterion indicated that the combination of these variables produced a multiple correlation of .93 and a corresponding R^2 of .86, signifying that the three variables accounted for 86 percent of the variance in the L2 reading performance. Comparisons between the predictions made by the Coh-Metrix L2 Reading Index and those made by the Flesch Reading Ease and Flesch-Kincaid Grade Level formulas indicated that the Coh-Metrix formula was significantly more accurate in predicting reading difficulty. The findings of Crossley et al. suggest that the incorporation of variables more closely aligned to psycholinguistic and cognitive reading processes improves the predictive ability of readability formulas and better assesses L2 text comprehensibility. However, one limitation of the study was its use of a corpus of strictly academic texts. Thus, the findings from the analysis could be generalized only to a specific genre of texts.

Method

Our purpose in this analysis is to compare the classification potential of traditional readability formulas (Flesch-Kincaid Grade Level and Flesch Reading Ease) to the Coh-Metrix L2 Reading Index in discriminating between levels of intuitively simplified, L2 reading texts (beginning, intermediate, and advanced). Our goal is to investigate which readability formula best classifies the text level. If intuitive text simplification results in the creation of texts that are easier to read because factors related to cohesion, decoding, parsing, and meaning construction have been manipulated, this should be reflected in the readability scores produced by the formulas. Thus, we expect that traditional readability formulas with their emphasis solely on word and sentence length, will not categorize levels of intuitively simplified texts with the same success as the Coh-Metrix L2 Reading Index, which is based on variables deemed important by psycholinguistic and cognitive accounts of reading.

To examine the classification potential of the readability formulas, we selected samples from a corpus of simplified news texts developed for L2 readers. The texts we sampled had not been manipulated based on standard readability formulas, but rather had been simplified based on intuitive notions. Using corpus, computational, and statistical approaches, this study examines the accuracy of the readability formulas in assigning each text to a specific level. We use the findings from this analysis to examine the construct validity of the readability formulas and how well they predict intuitive processes of text simplification used by material developers. We extend these findings into a general discussion about the processes of intuitive text simplification and their potential effects on text readability and comprehensibility.

Corpus Selection

The corpora used for this study is an extended version of the corpus used previously in Allen (2009). The texts which make up the corpus were taken from an English teaching website (www.onestopenglish.com). The website provides a popular and long-running service of offering simplified news texts and accompanying learning activities. The news texts were originally taken from the *Guardian Weekly*, a British-based publication with a wide international readership. The articles in the corpus were originally selected by the website editors for their topicality and interest value and typically center on world affairs. The texts are, therefore, non-academic in nature.

The news texts were simplified by a small, independent team of authors, into three levels of simplification: advanced, intermediate and beginning. Importantly, the method of simplification employed by the authors was intuitive, that is, without recourse to word lists, structural grading schemes, or readability formulas (Allen, 2009). However, as reported by Allen, the authors did provide certain indications regarding their approach to simplification. First, the authors followed the motto ‘grade the task, not the text,’ showing a tendency to only simplify when absolutely necessary. Second, a number of general strategies were employed, such as modifying idiomatic language at the intermediate level and removing it completely from the elementary level, while removing all passive structures and phrasal verbs from elementary level texts. Though the authors provided these mottos and strategies, it should be emphasized that they are not rules but are simply indications of the types of modifications made under an intuitive approach (Allen,

2009).

The total size of the news corpus used in this study is 210,538 words (*N* =300). The total sizes of the sub-corpora are as follows: Advanced = 76,579 words (*n* =100); Intermediate = 70,314 words (*n* = 100); Elementary = 63,645 (*n* =100). The sizes of the sub-corpora reflect the differences in text length due to the abridging of text at lower levels. However, text length differences are not a concern in this study because the Coh-Metrix indices found in the Coh-Metrix L2 Reading Index are normalized for text length. Descriptive statistics for the corpus is located in Table 1. We did not include the original, authentic texts in this analysis because our focus was on the process of simplification. Additionally, as reported by Allen (2009), the advanced texts are almost completely unmodified from the original texts.

Table 1. *Descriptive statistics for the simplified corpora*

	Mean number of words per text	*SD*	Mean number of paragraphs	*SD*
Beginner	636.450	162.239	9.820	2.935
Intermediate	703.140	164.082	10.110	2.934
Advanced	765.790	165.124	10.500	2.934

Selected Readability Formulas

To collect the readability of each text according to the various formulas, we used the computational tool Coh-Metrix (Graesser et al., 2004). The selected readability formulas are discussed below.

Flesch-Kincaid Grade Level. Coh-Metrix calculates the Flesch-Kincaid Grade Level based on the formula reported by Kincaid et al., (1975). The formula is based on the number of words per sentence (sentence length) and the number of syllables per word (word length). This formula is reported below.

Flesch-Kincaid Grade Level =
(0.39 x number of words/number of sentences)
+ (11.8 x number of syllables/number of words)
- 15.59

Flesch Reading Ease. Coh-Metrix calculates the Flesch Reading Ease based on the formula reported by Flesch (1948). Like Flesch-Kincaid Grade Level, this formula is based on the number of words per sentence (sentence length) and the number of syllables per word (word length). This formula is reported below.

Flesch Reading Ease =
206.835 - (1.015 x number of words/number of sentences)
- (84.600 x number of syllables/number of words)

Coh-Metrix L2 Reading Index. The Coh-Metrix L2 Reading Index is calculated using three linguistic indices reported by the Coh-Metrix tool. These three indices are CELEX Word

Frequency (logarithm mean for content words), Sentence Syntax Similarity (sentence to sentence adjacent mean), and Content Word Overlap (proportional adjacent sentences unweighted). These indices and their relation to text processing are discussed below.

Word Frequency. The CELEX Word Frequency index is based on frequency norms taken from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1993) a 17.9 million-word corpus. Word frequency effects have strong correlations with decoding in that frequent words are processed and understood more quickly than infrequent words (Haberlandt & Graesser, 1985; Just & Carpenter, 1980). In L2 reading studies, automatic decoding is argued to be an important predictor of reading performance (Koda, 2005).

Syntactic Parsing. The Sentence Syntax Similarity index measures the uniformity and consistency of parallel syntactic constructions in text both at the phrase level and the part of speech level. As a reader decodes a text, they assemble the decoded items into a syntactic structure (Just & Carpenter, 1980; Rayner & Pollatsek, 1994). If the syntactic structures are similar in construction, the cognitive demands on the reader are lower and more attention can be paid to meaning.

Word Overlap. The Content Word Overlap index measures how often content words overlap between two adjacent sentences. Content word overlap facilitates meaning construction and improves text comprehension and reading speed (Douglas, 1981; Kintsch & Van Dijk, 1978; Rashotte & Torgesen, 1985).

The Coh-Metrix L2 Reading Index as reported by Crossley et al. (2008) is below.

-45.032 + (52.230 x Content Word Overlap Value)
+ (61.306 x Sentence Syntax Similarity Value)
+ (22.205 x CELEX Frequency Value)

Statistical Analysis

We first conducted a series of repeated measures Analyses of Variance (ANOVA) to examine if all the readability formulas demonstrated significant differences between the levels of the reading texts. We then conducted a series of discriminant function analyses (DFAs) using each readability formula in turn in order to test the hypothesis that the readability formulas could differentiate between levels of simplified texts. DFAs are commonly used in analyses such as this one to distinguish text types (e.g., Biber 1993; Crossley & McNamara, 2009). In this study, the DFA was used to examine if the linguistic features contained within the formula were significant predictors of level classification.

Analysis

ANOVA

A series of repeated measures ANOVA was conducted using the selected readability formulas as the dependent variables and the simplified texts from the news corpus as the independent

variables. Descriptive statistics for the selected readability formulas are presented in Table 2. To control for Type 1 errors, we lowered our criterion for significance to $p = .015$. There was a significant linear trend for the Flesch-Kincaid Grade Level formula, F (2, 297) = 29.089, $p < .001$, $\eta p = .164$, the Flesch Reading Ease formula, F (2, 297) = 23.947, $p < .001$, $\eta p = .139$, and the Coh-Metrix Reading Index, F (2, 297) = 51.657, $p < .001$, $\eta p = .258$. Pairwise comparisons with Bonferroni corrections demonstrated that the Flesch Reading Ease formula and the Coh-Metrix L2 Reading Index yielded significant differences between each level of simplification. The Flesch-Kincaid Grade Level formula yielded significant differences between beginning and intermediate text readability scores, but not between intermediate and advanced text readability scores.

Table 2. *Means (Standard Deviations) for readability formulas and text levels.*

Variables	Beginner	Intermediate	Advanced
Flesch-Kincaid Grade Level	8.472 (1.613)	9.656 (1.703)	10.207 (1.612)
Flesch Reading Ease Score	63.978 (8.354)	58.806 (8.601)	55.506 (9.203)
Coh-Metrix L2 Reading Index	19.951 (4.151)	16.076 (5.312)	12.897 (5.198)

Accuracy of Model

To test the accuracy of the readability formulas to distinguish between the levels of L2 reading texts, we conducted a discriminant function analysis. We used the discriminant analysis to predict group membership (the level of the reading text) using a series of independent variables (the readability formulas). The DFA generates a discriminant function, which acts as the algorithm to predict group membership. First, the DFA is applied to the entire set. Later, the DFA from the entire set is used to predict group membership of the texts using repeated cross-validation. In cross-validation a fixed number of folds, or partitions of the data, is selected. Once the number of folds has been selected, each fold is used for testing the model. For this study, we selected a leave-one-out (n-fold) cross-validation model in which one instance in turn is left out and the remaining instances are used as the training set, in this case the 299 remaining texts. The accuracy of the model is tested on the model's ability to predict the omitted instance. This allows us to test the accuracy of the model on an independent data set. If the results of the discriminant analysis in both the entire set and the *n*-fold cross-validation set are similar, then the findings support the predictions of the analysis that readability formulas can be used to distinguish between simplified reading text levels. We report the findings of the DFA using an estimation of the accuracy of the analysis. This estimation is made by plotting the correspondence between the actual texts and the predictions made by the DFA model. We also report the results in terms of recall, precision, and F score. Recall scores are computed by tallying the number of hits over the number of hits + misses. Precision is the number of correct predictions divided by the number of incorrect predictions. We use both recall and precision because an algorithm could predict everything to be a member of a single group and score 100% in terms of recall. However, this could only happen at the expense of precision because the algorithm would have to claim members of the other group. By reporting both precision and recall, we can better understand the accuracy of the model. The F score can be thought of as a weighted average of the precision and recall results and is computed by dividing precision plus recall by precision multiplied by recall and multiplying the result by two.

Flesch-Kincaid Grade Level

The results demonstrate that the discriminant analysis for the Flesch-Kincaid Grade Level readability formula correctly allocated 148 of the 300 texts in the total set (df=4, *n*=300) $\chi 2$= 42.991, $p < .001$) for an accuracy of 49.3% (chance for this analysis is 33.3%). For the cross-validated set, the discriminant analysis correctly allocated 147 of the 300 texts for an accuracy of 49.0% (see Table 3 for results). The measure of agreement between the actual text type and that assigned by the model produced a Cohen's Kappa of 0.240, demonstrating a fair agreement.

Table 3. *Flesch-Kincaid Grade Level: Predicted Level versus Actual Level (total and cross-validated set)*

Actual text type	Predicted text type		
Total set	Beginner	Intermediate	Advanced
Beginner	65	12	23
Intermediate	33	22	45
Advanced	24	15	61
Cross-validated set	Beginner	Intermediate	Advanced
Beginner	65	12	23
Intermediate	34	21	45
Advanced	24	15	61

The precision and recall scores from the model for predicting the level of the simplified reading texts can be found in Table 4. In terms of recall, the model performed best at recalling beginner texts (classifying 123 texts as beginner level of which 65 were beginner level) and performed worst at recalling intermediate texts (classifying 48 texts as intermediate level of which 21 were intermediate level). In terms of precision, the model was best at predicting beginner level texts (correctly classifying 65 of the 100 beginner level texts) and worst at predicting intermediate level texts (classifying 21 of the 100 intermediate texts). The overall accuracy of the model for the total set was .488 (the average F score). The accuracy for the cross-validated set was .484.

Table 4. *Precision and recall results for Flesch-Kincaid Grade Level (total and cross-validated set)*

Total set	Recall	Precision	F1
Beginner	0.533	0.650	0.591
Intermediate	0.449	0.220	0.334
Advanced	0.473	0.610	0.541
Cross-validated set	Recall	Precision	F1
Beginner	0.528	0.650	0.589
Intermediate	0.438	0.210	0.324
Advanced	0.473	0.610	0.541

Flesch Reading Ease

The results demonstrate that the discriminant analysis for the Flesch Reading Ease readability formula correctly allocated 133 of the 300 texts in the total set (df=4, *n*=300, $\chi 2$= 28.238, p < .001) for an accuracy of 44.3% (chance for this analysis is 33.3%). For the cross-validated set, the discriminant analysis also correctly allocated 133 of the 300 texts for an accuracy of 44.3% (see Table 5 for results). The measure of agreement between the actual text type and that assigned by the model produced a Cohen's Kappa of 0.165, demonstrating a slight agreement.

Table 5. *Flesch Reading Ease Score: Predicted level versus actual level (total and cross-validated set)*

Actual text type	Predicted text type		
Total set	Beginner	Intermediate	Advanced
Beginner	63	13	24
Intermediate	38	18	44
Advanced	27	21	52
Cross-validated set	Beginner	Intermediate	Advanced
Beginner	63	13	24
Intermediate	38	18	44
Advanced	27	21	52

The precision and recall scores from the model for predicting the level of the simplified reading texts can be found in Table 6. For recall, the model performed best at recalling beginner texts (classifying 128 texts as beginner level of which 63 were beginner level) and performed worst at recalling intermediate texts (classifying 52 texts as intermediate level of which 18 were intermediate level). In reference to precision, the model was best at predicting beginner level texts (correctly classifying 63 of the 100 beginner level texts) and worst at predicting intermediate level texts (classifying 18 of the 100 intermediate texts). The overall accuracy of the model for the total set was .436 (the average F score). The accuracy for the cross-validated set was also .436.

Table 6. *Precision and recall results for Flesch Reading Ease Score (total and cross-validated set)*

Total set	Recall	Precision	F1
Beginner	0.508	0.630	0.569
Intermediate	0.346	0.180	0.263
Advanced	0.433	0.520	0.477
Cross-validated set	Recall	Precision	F1
Beginner	0.508	0.630	0.569
Intermediate	0.346	0.180	0.263
Advanced	0.433	0.520	0.477

Coh-Metrix L2 Reading Index

The results demonstrate that the discriminant analysis for the Coh-Metrix L2 Reading Index readability formula correctly allocated 180 of the 300 texts in the total set (df =4, n=300) $\chi 2$= 118.487, $p < .001$) for an accuracy of 60.0% (chance for this analysis is 33.3%). For the cross-validated set, the discriminant analysis also correctly allocated 180 of the 300 texts for an accuracy of 60.0% (see Table 7 for results). The measure of agreement between the actual text type and that assigned by the model produced a Cohen's Kappa of 0.400, demonstrating a moderate agreement.

Table 7. *Coh-Metrix L2 Reading Index: Predicted level versus actual level (total and cross-validated set)*

Actual text type	Predicted text type		
Total set	Beginner	Intermediate	Advanced
Beginner	70	21	9
Intermediate	27	39	34
Advanced	7	22	71
Cross-validated set	Beginner	Intermediate	Advanced
Beginner	70	21	9
Intermediate	27	39	34
Advanced	7	22	71

The precision and recall scores from the model for predicting the level of the simplified reading texts can be found in Table 8. In reference to recall, the model performed best at recalling beginner texts (classifying 104 texts as beginner level of which 70 were beginner level) and performed worst at recalling intermediate texts (classifying 82 texts as intermediate level of which 39 were intermediate level). In reference to precision, the model was best at predicting advanced level texts (correctly classifying 71 of the 100 advanced level texts) and worst at predicting intermediate level texts (classifying 39 of the 100 intermediate texts). The overall accuracy of the model for the total set was .587 (the average F score). The accuracy for the cross-validated set was .587.

Table 8. *Precision and recall results for Coh-Metrix L2 Reading Index (total and cross-validated set)*

Total set	Recall	Precision	F1
Beginner	0.625	0.700	0.663
Intermediate	0.476	0.390	0.433
Advanced	0.623	0.710	0.666
Cross-validated set	Recall	Precision	F1
Beginner	0.625	0.700	0.663
Intermediate	0.476	0.390	0.433
Advanced	0.623	0.710	0.666

Comparisons between Formulas

We assigned each text either a 0 or a 1 based on whether the readability formula accurately predicted its group membership (0 = inaccurate, 1 = accurate). We then conducted *t* tests between the classification results for each readability formula to examine if significant differences in classification accuracy existed between the formulas. As in our ANOVA analysis, we lowered our criterion for significance to $p = .015$ to control for Type 1 errors. No significant differences in classification accuracy were noted between the two traditional readability formulas $t(598) = 1.227$, $p > .015$. Significant differences in classification accuracy were reported between the Coh-Metrix L2 Reading Index and Flesch-Kincaid Grade Level $t(598) = -2.635$, $p < .010$ and the Coh-Metrix L2 Reading Index and Flesch Reading Ease $t(598) = -3.883$, $p < .001$. The results demonstrate that the predictions made by the Coh-Metrix L2 Reading Index were significantly more accurate than those made by the traditional readability formulas. The mean scores and standard deviations for this analysis are located in Table 9. Perfect predictive ability would be reflected by a mean score of one.

Table 9. *Descriptive statistics for t-test data*

Readability formula	Mean	*SD*	N
Flesch-Kincaid Grade Level	0.493	0.501	300
Flesch Reading Ease	0.443	0.498	300
Coh-Metrix Reading Index	0.600	0.491	300

Discussion

This study has demonstrated that a readability formula based on psycholinguistic and cognitive models of reading and traditional readability formulas can significantly classify texts based on their levels of intuitive text simplification. However, the accuracy scores are significantly higher for the Coh-Metrix L2 Reading Index, indicating that the Coh-Metrix index is better able to discriminate between the different levels of texts. This finding is backed by the reported effect sizes. In total, these findings support the notion that the variables used in the Coh-Metrix L2 Reading Index are more closely aligned to the intuitive text processing used by L2 material writers when simplifying reading texts than those variables provided by traditional readability formulas. These findings also provide some evidence for the manner in which the discourse processes used in intuitive text simplification modify the linguistic construction of text.

We begin our discussion with a comparison of the categorical accuracies reported by the targeted readability formulas. We expected the Coh-Metrix L2 Reading Index to outperform the traditional readability indices because we hypothesized that the indices included in the Coh-Metric L2 Reading Index had stronger conceptual overlap to variables featured in psycholinguistic and cognitive accounts of reading. This hypothesis presupposes that intuitive text simplification would reflect text modifications based on these same psycholinguistic and cognitive accounts. These predictions were supported in the statistical analysis, which demonstrated that traditional readability formulas were weaker classifiers of reading text level than the Coh-Metrix L2 Reading Index. The weakest classifier was Flesch Reading Ease, which

correctly classified 44% of the texts into their respective levels. The Kappa value of .165 demonstrated only a slight agreement, which shows that the strength of the relationship between the classification ability of the formula and the actual level classification was weak. The Flesch Reading Ease Score was particularly weak at classifying intermediate texts (18% accuracy). The Flesch-Kincaid Grade Level formula performed better than the Flesch Reading Ease Formula (but not significantly better) and correctly classified 48% of the texts based on grade level. The reported Kappa value between the actual text classification and the classification made by the formula was *.240*, demonstrating that the strength of the relationship between the two was fair. As with the Flesch Reading Ease formula, the Flesch-Kincaid Grade Level formula had the most difficulty classifying intermediate texts (22%). The best predictor of level classification was the Coh-Metrix L2 Reading Index, which correctly classified 59% of the reading texts by level. The reported Kappa value of *.400* demonstrated a moderate relationship between the predictions made by the formula and the actual classifications. As with the traditional readability formulas, the Coh-Metrix L2 Reading Index performed least well in classifying intermediate texts (39% accuracy). The difficulty for all formulas in classifying the intermediate level texts likely resulted from the transitory nature of the level as a product of the simplification process (Allen, 2009). The authors of the simplified texts work first at simplifying the advanced text down to intermediate level then use the resulting intermediate text as the basis for simplifying further to elementary. Thus, as a transitory level, the intermediate texts share similarities with both beginning and advanced texts, many of which are not shared between beginning and advanced texts. These similarities likely produce misclassifications on the part of the formulas because the features attributed to both beginning and advanced texts are found at the intermediate level.

A more interesting story lies in the discussion of the strengths of the tested readability formulas to assess texts that have been intuitively simplified. The weaker classification results for the traditional readability formulas underscore the criticism they have received in the past. This criticism includes arguments that traditional readability formulas are only weakly related to cognitive processes important in reading and do not consider text comprehension factors. Our findings support the notion that writers engaged in intuitive text simplification do not simply select words that are shorter or reduce the number of words in sentences, but instead focus more on features related to text comprehensibility and cognitive reading processes. Evidence for this finding is found in the increased accuracy of the Coh-Metrix L2 Reading Index and the failure of the Flesch-Kincaid Reading Level formula to report significant differences between intermediate and advanced texts. The improvements in level classification that came from using a readability formula that includes indices of text comprehensibility (cohesion and meaning construction) and indices more closely aligned with the cognitive processes of reading (decoding and syntactic parsing) provide evidence that the process of intuitive simplification involves such factors. Assuming that such features are related to reading processes and that attention to such features in the simplification process will lead to more comprehensible texts, we argue for the benefits of intuitive text simplification over text simplification using traditional readability formulas. At the same time, we argue that the simplistic mechanisms found in traditional readability formulas are less likely to capture text features related to text comprehensibility and cognitive processing, thus highlighting a primary weakness of traditional readability formulas that is well documented in past research (e.g., Carrell, 1987; Crossley, Greenfield, & McNamara, 2008; Davison & Kantor, 1982). We do note, of course, that the traditional readability formulas classified texts into appropriate categories at a level above chance. Thus, to some degree, traditional readability

formulas do to measure levels of text difficulty related to cognitive reading processes such as decoding and syntactic parsing. However, traditional readability formulas are less precise than the Coh-Metrix L2 Reading Index, underscoring the need to develop more valid and accurate readability formulas.

This study also provides some evidence for the extendibility of the Coh-Metrix L2 Reading Index to other genres such as news articles. One concern with traditional readability formulas is that whereas they may work reasonably well for strictly academic genres, their use has been broadened to include many genres that are beyond the formulas' original purposes (Greenfield, 1999). This study demonstrates that the Coh-Metrix L2 Reading Index had moderate degrees of success at classifying news texts and was significantly more accurate than traditional readability formulas. Such a finding supports the notion that the Coh-Metrix L2 Reading Index may be extendible to genres outside of strictly academic texts. One reservation for this claim, though, is that academic and news texts appear to share many common linguistic features (Biber, Johannson, Leech, Conrad, & Finegan, 1999) such as the high frequency of nouns versus pronouns, repetition of nouns with definite articles, and the constitution of complex noun phrases. Nevertheless, the increased accuracy of the Coh-Metrix L2 Reading Index outside of traditional academic genres bodes well for the potential extendibility of the formula.

Conclusion

Overall, this study has demonstrated the benefits of the Coh-Metrix L2 Reading Index in classifying and examining differing levels of intuitively simplified texts over traditional readability formulas. The study also finds support for intuitive simplification processes in that they appear to follow principles important in psycholinguistic accounts of text processing and cognitive accounts of text comprehensibility. This is especially true for the texts analyzed in this study.

While the performance of the Coh-Metrix L2 Reading Index surpassed that of traditional readability formulas, it must be noted that it only classified 59% of the text levels accurately and demonstrated only a moderate agreement with the actual level classification. Much of this low classification rate can be attributed to the intermediate level texts, which were difficult for all the readability formulas to classify. However, it is also likely that many of the intuitive simplification features in the texts that lead to better text comprehension were not measured by the Coh-Metrix L2 Reading Index. Such an assumption rests on the notion that the reading index only considers three variables, while the process of intuitive text simplification likely modifies a much larger number of linguistic features. Such an assumption does not challenge the strength of the Coh-Metrix L2 Reading Index, especially when compared to traditional readability formulas, but it does suggest that more research is needed to develop formulas that contain more linguistic features and that better match text readability for various genres, readers, and levels. Specifically, as noted by Crossley et al. (2008), larger reading studies need to be conducted to improve the formula and allow for the inclusion of additional variables. These reading studies should include L2 learners at various proficiency levels and from various first language backgrounds. Additionally, the criteria used in such studies should include both authentic and simplified texts. These simplified texts should be controlled by approaches so that learners read

texts simplified intuitively, by structure and word lists, and by readability formulas. In this way it will be possible to further assess the validity of advanced readability formulas for predicting text comprehensibility.

Acknowledgments

This research was supported by a grant from the Institute of Education Sciences (IES: R3056020018-02) to the University of Memphis. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of IES. Correspondence concerning this article should be addressed to the first author.

References

Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, *37* (4), 585–599.

Allen, J., & Widdowson, H. G. (1979). Teaching the communicative use of English. In C. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 124–142). Oxford: Oxford University Press.

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

Bamford, J. (1984). Extensive Reading by Means of Graded Readers. *Reading in a Foreign Language*, *2* (2), 218–260.

Biber, D. (1993). Representativeness in corpus design. *Journal of Literary and Linguistic Computing*, *8* (4), 243–257.

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

Blau, E. K. (1982). The effect of syntax on readability for ESL students in Puerto Rico. *TESOL Quarterly*, *16*, 517–528.

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*, *20*, 7–36.

Carrell, P. L. (1987) Readability in ESL. *Reading in a Foreign Language*, *4* (1), 21– 40.

Crossley, S. A., Allen, D., & McNamara, D. S. (in press). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research.*

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 and 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. M., McCarthy, P. M. & McNamara, D. S. (2007). A Linguistic Analysis of Simplified and Authentic Texts. *The Modern Language Journal*, *91* (1), 15–30.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using

cognitively based indices. *TESOL Quarterly*, *42*(3), 475–493.
Crossley, S. A. & McNamara, D. S. (2008). Assessing Second Language Reading Texts at the Intermediate Level: An approximate replication of Crossley, Louwerse, McCarthy, and McNamara (2007). *Language Teaching*, *41* (3), 409–229.
Crossley, S. A. & McNamara, D. S. (2009). Computationally assessing lexical differences in second language writing. *Journal of Second Language Writing*, *17*(2), 119–135.
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 Center for Applied Linguistics.
Flesch, R. (1948). A new readability yardstick. *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* (3–22). Mahwah New Jersey: Lawrence Erlbaum Associates.
Goodman, K., & Freeman, D. (1993). What's Simple in Simplified Language. In M .L. Tickoo (Ed.), *Simplification: Theory and application* (pp. 69–76). Singapore: SEAMEO Regional Language Center.
Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). *Coh-Metrix*: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and 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 Experiment Psychology: General*, *114*, 357–374.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354.
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.
Kuo, C. (1993). Problematic issues in EST materials development. *English for Specific Purposes*, *12*, 171–181.
Long, M., & Ross, S. (1993). Modifications that Preserve Language and Content. In M. L. Tickoo (Ed.), *Simplification: Theory and application* (pp. 29–52). Singapore: SEAMEO Regional Language Center.
Lotherington-Woloszyn, H. (1993). Do Simplified Texts Simplify Language Comprehension for ESL Learners? In M. L. Tickoo (Ed.), *Simplification: Theory and application* (pp.140–154). Singapore: SEAMEO Regional Language Center.
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., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In J. D. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 60–81). Mahwah, NJ: Erlbaum.
Onestopenglish, 2007. News Lessons [online]. Macmillan English Campus, London. Available from: http://www.onestopenglish.com (accessed 28.08.2007).
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, New Jersey: Prentice Hall.
Simensen, A.M. (1987). Adapted readers: How are they adapted? *Reading in a Foreign Language*, *4*, 41–57.
Tweissi, A. I. (1998). The effects of the amount and the type of simplification on foreign language reading comprehension. *Reading in a Foreign Language*, *11*, 191–206.
Yano, Y, Long, M., & Ross, S. (1994). Effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, *44* (2), 189–219.
Young, D. J. (1999). Linguistic Simplification of Second Language Reading Material: Effective Instructional Practice? *The Modern Language Journal*, *83* (3), 350–366.

About the Authors

Scott Crossley is an Assistant Professor Georgia State University. His interests include computational linguistics, corpus linguistics, discourse processing, and discourse analysis. He has published articles in genre analysis, multi-dimensional analysis, discourse processing, speech act classification, cognitive science, and text linguistics. Email: sacrossley@gmail.com

David Allen is a Project Assistant Professor at the University of Tokyo. He is interested in psycholinguistics, corpus linguistics, second language reading and writing, and genre analysis. He is currently pursuing a doctorate focusing on the representation and processing of cognates by Japanese-English bilinguals. Email: dallen@aless.c.u-tokyo.ac.jp

Danielle McNamara is a Professor at the University of Memphis and Director of the Institute for Intelligent Systems. 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. Email: dsmcnamara1@gmail.com