Please ReaderBench This Text

In: Tutoring and Intelligent Tutoring Systems
Editor: Scotty D. Craig
ISBN: 978-1-53614-085-9

Chapter 9

PLEASE *READERBENCH* THIS TEXT: A MULTI-DIMENSIONAL TEXTUAL COMPLEXITY ASSESSMENT FRAMEWORK

Mihai Dascalu[1,2,*], Scott A. Crossley[3], Danielle S. McNamara[4], Philippe Dessus[5] and Stefan Trausan-Matu[1,2]

[1]Computer Science Department, University Politehnica of Bucharest, Bucharest, Romania
[2]Academy of Romanian Scientists, Bucharest, Romania
[3]Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA, US
[4]Institute for the Science of Teaching and Learning, Arizona State University, Tempe, AZ, US
[5]University Grenoble Alpes, LaRAC, Grenoble, France

ABSTRACT

A critical task for tutors is to provide learners with suitable reading materials in terms of difficulty. The challenge of this endeavor is increased by students' individual variability and the multiple levels in which complexity can vary, thus arguing for the necessity of automated systems to support teachers. This chapter describes *ReaderBench*, an open-source multi-dimensional and multi-lingual system that uses advanced Natural Language Processing techniques to assess textual complexity at multiple levels including surface-based, syntax, semantics and discourse structure. In contrast to other existing approaches, *ReaderBench* is centered on cohesion and makes extensive usage of two complementary models, i.e., Cohesion Network Analysis and the polyphonic model inspired from dialogism. The first model provides an in-depth view of discourse in terms of cohesive links, whereas the second one highlights interactions between points of view

* Corresponding Author: mihai.dascalu@cs.pub.ro

spanning throughout the discourse. In order to argue for its wide applicability and extensibility, two studies are introduced. The first study investigates the degree to which *ReaderBench* textual complexity indices differentiate between high and low cohesion texts. The *ReaderBench* indices led to a higher classification accuracy than those included in prior studies using *Coh-Metrix* and *TAACO*. In the second study, *ReaderBench* indices are used to predict the difficulty of a set of various texts. Although the high number of predictive indices (50 plus) accounted for less variance than previous studies, they make valuable contributions to our understanding of text due to their wide coverage.

Keywords: comprehension modeling, learning analytics, automated essay scoring, data analytics, Natural Language Processing

INTRODUCTION

Intelligent Tutoring Systems have been developed to support learners and human tutors by providing customized instructions and feedback. Measuring text difficulty is of particular interest in learning activities in order to best match resources to reader's comprehension level. However, the automated assessment of text difficulty is a difficult endeavor because multiple text features and learner characteristics (e.g., prior knowledge, language familiarity, reading performance, cognitive capabilities) need to be taken into account in order to provide tailored feedback. Thus, automated systems need to be adaptive to the audience, ensuring that learners are challenged, but are not excessively frustrated or demotivated.

Several automated systems have been developed and adopted in a wide range of educational scenarios. *E-Rater* (Powers, Burstein, Chodorow, Fowles, & Kukich, 2001) is one of the first systems to automatically evaluate essays in which the quality of writing was measured by extracting a set of text indices such as: lexical complexity, spelling mistakes, lexical diction, text organization, as well as basic content analyses based on vocabularies. *E-Rater* supports a multi-layered textual complexity evaluation and subsequent software releases of the system added new features, including a model for evaluating the complexity of inferences in the discourse (Grosz, Weinstein, & Joshi, 1995). Newer tools such as TAALES (Kyle, Crossley, & Berger, in press), TAACO (Crossley, Kyle, & McNamara, 2016), TAASC (Kyle & Crossley, 2018), or *Coh-Metrix* (Graesser, McNamara, Louwerse, & Cai, 2004) provide comprehensive lists of scores for specific textual complexity indices that can be used for a wide range of text analyses on cohesion (i.e., semantic relations that exist and define a text) and language. Starting from automated essay scoring, the aim has transcended towards building Intelligent Tutoring Systems that provide detailed feedback besides singular evaluation scores. Automated Writing Evaluation systems such as the Writing Pal (Roscoe, Varner, Weston, Crossley,

& McNamara, 2014) take the assessment component one step further and provide feedback to learners, thus supporting them to improve their writing skill.

Our implemented framework, *ReaderBench* (Dascalu, 2014; Dascalu, Dessus, Bianco, Trausan-Matu, & Nardy, 2014; Dascalu et al., 2017) integrates an extensive list of textual complexity indices centered on cohesion and discourse structure. *ReaderBench* is built on top of Cohesion Network Analysis (CNA; Dascalu, McNamara, Trausan-Matu, & Allen, 2018) which provides an in-depth longitudinal perspective over the cohesive links across the text. Moreover, *ReaderBench* also integrates a complementary transversal perspective (Dascalu, Trausan-Matu, McNamara, & Dessus, 2015), namely the polyphonic model of discourse (Trausan-Matu, Stahl, & Sarmiento, 2007), which highlights interactions between points of view (i.e., "voices"). Thus, *ReaderBench* introduces in-depth textual complexity indices that account for certain specificities of the comprehension process, namely: a) automated word learning curves in our Age of Exposure model (Dascalu, McNamara, Crossley, & Trausan-Matu, 2015); b) document cohesion flow metrics that model the structure of discourse in terms of global cohesion (Crossley, Dascalu, Trausan-Matu, Allen, & McNamara, 2016); c) metrics inspired from dialogism (Dascalu, Allen, McNamara, Trausan-Matu, & Crossley, 2017) that consider any text as a dialogue in which multiple points of view or voices interact and inter-animate; and d) rhythm features inspired from prosody (Balint, Dascalu, & Trausan-Matu, 2016b).

The multi-layered textual complexity model behind *ReaderBench* is highly extensible and supports different languages, such as: *English* (Allen, Dascalu, McNamara, Crossley, & Trausan-Matu, 2016; Crossley, Paquette, Dascalu, McNamara, & Baker, 2016; Dascalu, Popescu, Becheru, Crossley, & Trausan-Matu, 2016), *French* (Dascalu, Dessus, Bianco, & Trausan-Matu, 2014; Dascalu, Stavarache, Trausan-Matu, Dessus, & Bianco, 2014), *Romanian* (Dascalu, Gifu, & Trausan-Matu, 2016; Gifu, Dascalu, Trausan-Matu, & Allen, 2016), and *Dutch* (Dascalu, Westera, Ruseti, Trausan-Matu, & Kurvers, 2017). In addition, several languages including Spanish, Italian, and Latin are partially covered.

At present, *ReaderBench* has more than a thousand textual indices for English language that cover: classic readability formulas that serve as baseline measures, surface indices (e.g., character, word, sentence and paragraph counts), syntax measures, and, more importantly, semantics and discourse structure indices to highlight in-depth comprehension processes. All the previous categories of indices are included in a multi-layered and multi-lingual textual complexity model, and are described in detail in the following section. Thus, the aim of *ReaderBench* is to ensure a high degree of flexibility to adapt to different learning scenarios, playing the role of an artificial tutor that can, upon request at any moment, gauge multiple features of a text. To this purpose, our framework has been subject to extensive validations and this book chapter is focused on two specific studies, each centered on a different task. The first study considers perceived text difficulty as a function of cohesion in multi-paragraph texts. The second study is

centered on predicting text difficulty by replicating the findings for the readability scores from previous studies (Crossley, Skalicky, Dascalu, Kyle, & McNamara, 2017) which used different tools.

READERBENCH – A COMPREHENSIVE MULTI-LAYERED AND MULTI-LINGUAL TEXTUAL COMPLEXITY MODEL

Figure 1 depicts an overview of our multi-layered textual complexity model which includes: a) the simplest *surface* measures that account only for the form of the text; b) *syntactic* indices computed at sentence level which consider the distribution of different parts of speech or of syntactic dependencies; c) *semantics* and *complex discourse structure* which are mostly derived from our CNA model; and d) *word* complexity focused on individual tokens and spanning across all the previous levels by including different facets of the difficulty of each word taken individually, or within its semantic context.

The vast majority of the aforementioned indices are language independent, and thus only specific semantic models need to be trained once the Natural Language Processing (NLP) pipeline (Manning & Schütze, 1999) is in place, whereas some indices are language-specific and require additional NLP techniques to be set up. *ReaderBench* includes a comprehensive NLP pipeline (Manning & Schütze, 1999) that considers: a) stop-word elimination, b) the reduction of inflected forms to their corresponding lemmas, c) named entity recognition, d) the annotation of each word with its corresponding part of speech (POS) tag, e) dependency parsing, and f) co-reference resolution. This chapter is focused on providing an overview of the English indices, which represent the largest subset of available textual complexity indices and are used in the current studies.

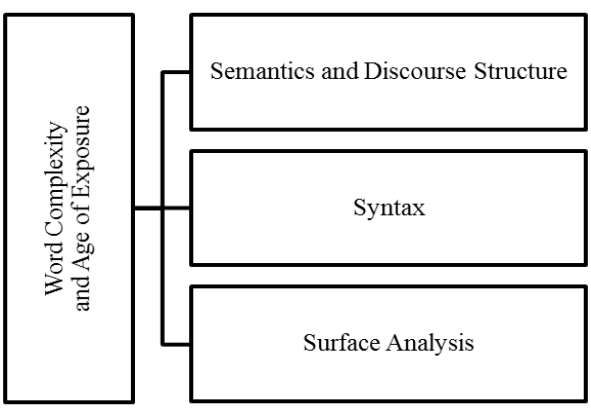

Figure 1. *ReaderBench* multi-layered textual complexity model.

Surface Analysis

Surface analysis considers basic traits of texts derived from the initial studies of Page (1966, 1968) and Wresch (1993) that include statistically and easily computable *proxes* (i.e., computer automated approximations of interest) like: a) paragraph and sentence lengths (average values and standard deviations) in terms of characters, words or unique words; b) commas per sentence or paragraph; c) entropy measures (Shannon, 1948, 1951) at character and word levels. Character entropy is in general a language feature, i.e., it has similar values between texts written in the same language based on the distribution of individual letter. In contrast, higher entropy computed on the distribution of word stems that capture the common root of related concepts, reflects a more complex text because more varied concepts are introduced.

Syntax

Word- and sentence-based analyses, including part of speech tagging and dependency parsing, play important roles by providing two different complexity assessment schemes: a) normalized frequencies of each part of speech and b) structural indices derived from the parsing tree. Although nouns and verbs are the most common POSs, our analysis is particularly aimed at prepositions, adjectives, and adverbs that are potentially indicative of a more elaborate and complex text structure. Moreover, pronouns are indicative of a more inter-twined and linked structure of the discourse by revealing potential pronominal co-references. In addition, multiple indices can be derived using the structure of the parsing tree (e.g., an increased number of specific semantic dependencies or a higher maximum depth indicate a more complex discourse structure, yielding increased textual complexity) (Gervasi & Ambriola, 2002).

Entity-density features are included because the number of entities within a text impacts the cognitive resources needed for their understanding, which impacts text readability. In general, named entities introduce conceptual information required for contextualizing the text; thus, in order to quantify the difficulty of a text with regard to newly introduced concepts in a text, we compute counts of entities (unique or not) per paragraph or sentence, as well as the percentages of named entities that are also nouns.

Semantics, Cohesion Network Analysis, and Discourse Structure

Text *cohesion* relates to explicit lexical, grammatical, or semantic text cues that support readers in making connections among text segments and the underlying ideas. Cohesion characterizes a unified and connected text, with sentences and paragraphs

related to one another using explicit cues in the text (McNamara, Graesser, & Louwerse, 2012). Cohesion relates to humans' perception of that text's overall quality and coherence, and may be present at both local (i.e., sequential relations between neighboring sentences) and global (i.e., relations between paragraphs) levels (Crossley, Roscoe, McNamara, & Graesser, 2011; McNamara, Crossley, & McCarthy, 2010).

ReaderBench makes extensive use of Cohesion Network Analysis (Dascalu, McNamara, Trausan-Matu, & Allen, 2018), which provides an in-depth view of the cohesive links that connect the discourse. *Cohesion* is viewed from a computational perspective as a relatedness measure between text chunks computed using multiple semantic models that complement one another (Dascalu, 2014). Within *ReaderBench*, cohesion is the average score between the Wu-Palmer semantic distance in WordNet (Wu & Palmer, 1994) and semantic similarities measured using the following models: Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003), and word2vec (Mikolov, Chen, Corrado, & Dean, 2013). Therefore, local and global cohesion are reflected within CNA as the strength of intra- and inter-paragraph links extracted from the cohesion graph.

Crossley, Dascalu, Trausan-Matu, Allen, and McNamara (2016) developed automated measures of global document cohesion flow based on a CNA that considers the adequacy of paragraph sequences – i.e., the degree to which one paragraph succeeds the previous one in a cohesive manner. Our *cohesion flow* measures consider the order of different paragraphs and the manner in which they are combined to hold the document together. A text that exhibits a high cohesion flow by linking ideas between adjacent paragraphs tends to be easier to comprehend.

In addition, *ReaderBench* also implements the polyphonic model (Trausan-Matu, Stahl, & Sarmiento, 2007) inspired from *dialogism* in which the discourse is perceived as an inter-animation of different points of view (i.e., "voices") that interact with each other. After identifying voices as semantic chains of related words spanning throughout the text (Dascalu, 2014), several textual complexity indices are computed in order to quantify the impact of each voice and to establish the degree of overlap between voices (Dascalu, Allen, McNamara, Trausan-Matu, & Crossley, 2017): a) distribution per sentence or paragraph, including span (distance between the last and the first occurrence of words from the same voice), b) recurrence (average and standard deviation of the distance between subsequent words pertaining to the same voice), and c) overlap measures (e.g., co-occurrences or mutual information).

Related to discourse, *ReaderBench* accounts for three additional dimensions. First, *pronominal resolutions* are performed (Lee et al., 2011; Manning et al., 2014) and several complexity indices are computed including: a) the average number of co-references per chain, their span and inference distance, b) the average number of active co-reference chains per word (if more words are included in co-reference chains, the text becomes harder to comprehend as it is more inter-twined and more inferences need to be resolved),

and c) the number of co-reference chains with a large span (experimentally set at 30% of the document length).

Second, *rhythm* is an important feature of discourse (Trausan-Matu, Dascalu, & Rebedea, 2014). Rhythm in a text is related to its communicative purposes and genre (Balint, Dascalu, & Trausan-Matu, 2016a, 2016b). Similar to previous dimensions, multiple indices are integrated into *ReaderBench*, namely: a) the average number of stressed syllables and of rhythmic units in each sentence, b) the number of deviations from dominant structures divided by total number of syllabic segments, c) the rhythmic index (Marcus, 1970), d) the frequency of the maximum rhythmic index, e) the maximum number of consecutive unstressed syllables, and f) the number of alliterations and assonances in a text searching sentence by sentence.

Third, *cue phrases* are used to quantify the usage of different types of pronouns (e.g., first, second, third, interrogative, or indefinite) and connectives (e.g., conjuctions, contrasts, sentece linking, or conditions), thus providing insights to the structure of each sentence or paragraph, and its corresponding degree of complexity. Finally, specific word lists that capture particular semantic valences are integrated into *ReaderBench*, namely: *General Inquirer* (GI, http://www.wjh.harvard.edu/~inquirer/homecat.htm) (Stone, Dunphy, Smith, Ogilvie, & associates, 1966), *Lasswell* (Lasswell & Namenwirth, 1969) dictionary, SenticNet (Cambria, Grassi, Poria, & Hussain, 2013), Affective Norms for English Words (ANEW) (M. M. Bradley & Lang, 1999), Geneva Affect Label Coder (GALC) (Scherer, 2005), Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Booth, & Francis, 2007), and *EmoLex* or *NRC* Word-Emotion Association Lexicon (Mohammad & Turney, 2013). The cue phrases and taxanomies represent a solid background for covering the most representative linguistic approaches relying on word counts.

Word Complexity and Age of Exposure

Word complexity is computed as a mix of multiple word indices, such as: a) syllable count, b) length of suffices and prefixes expressed in number of characters between a word's inflected form and its lemma or stem, c) specificity computed as the inverse document frequency from the training corpora, d) word polysemy count and average/maximum hypernym tree depth computed using WordNet, and e) our metric of word difficulty that considers contextualization – Age of Exposure (AoE) (Dascalu, McNamara, Crossley, & Trausan-Matu, 2015). Similar to Landauer, Kireyev, and Panaccione (2011), our goal was to create a word learning model (AoE) using LDA, an automated alternative for Age of Acquisition (AoA) scores – e.g., Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012), Bird, Franklin, and Howard (2001) or

Schock, Cortese, Khanna, and Toppi (2012) that estimate a learner's age at which a certain word is correctly understood.

In addition, *ReaderBench* considers the following individual word complexity measures: a) *mean syllable count per word:* longer, more complex words tend to be perceived as being more difficult; b) *mean polysemy count per word:* words with multiple senses are more difficult and require more contextual information for disambiguation; c) average and maximum *distance* within the *hypernym tree to the ontology root:* more general words are closer to the root, whereas more specific words tend to have a longer path; and d) *differences* between the *inflected form,* the *lemma* and the *stem:* words with longer prefixes and suffixes tend to be more complex. All word indices are averaged at document, paragraph and sentence levels by considering only the lemmas of content words (i.e., dictionary words, not included in the stop-words lists, and having as part of speech one of the following: noun, verb, adverb or adjective).

PREDICTING COMPREHENSION THROUGH AUTOMATED TEXT ANALYSIS

The following studies represent in-depth comprehension analyses performed using the previously presented *ReaderBench* textual complexity indices for English language, and are centered on the following research questions:

1) What automated text complexity indices best capture differences in text cohesion?
2) Are linguistic features predictive of judgments of text comprehension, processing, and familiarity?

Study 1 – Text Cohesion

Corpus

We first selected the texts compiled by McNamara, Louwerse, McCarthy, and Graesser (2010). These texts were used in experiments that investigated multi-paragraph text cohesion. The dataset contains 19 pairs of texts (low and high cohesion texts) from 12 different studies discourse studies. The texts (n = 38) were selected from textbooks, encyclopedia articles, researcher created texts, or text from books. In each case, the original text was manipulated by discourse processing experts to make the text more cohesive. In this study, we used *ReaderBench* indices to discriminate between the original texts and those texts modified to make them more cohesive.

Statistical Analyses

A number of pre-selection criteria were required for *ReaderBench* indices to be included in the final analysis. First, *ReaderBench* indices that yielded non-normal distributions were removed because they violated statistical assumptions. Of the original 329 indices, this step removed 117 of the indices. A multivariate analysis of variance (MANOVA) was then conducted to examine which *ReaderBench* variables reported meaningful and significant differences between the original texts and the text modified by experts to make them more cohesive. The MANOVA was followed by a stepwise discriminant function analysis (DFA) using the selected *ReaderBench* indices that demonstrated significant differences between the original and modified texts. Those indices that demonstrated significant differences between the two text groups and did not exhibit multicollinearity ($r > .70$) with other indices in the set were used in the DFA. In the case of multicollinearity, the index demonstrating the largest effect size was retained in the analysis. The DFA was used to develop an algorithm to predict group membership through a discriminant function co-efficient. A DFA model was first developed for all of the original and modified texts. This model was then used to predict group membership of the original and modified texts using leave-one-out-cross-validation (LOOCV) in order to ensure that the model was stable across the dataset.

Results

MANOVA

A MANOVA was conducted using *ReaderBench* indices as the dependent variables and the original and modified texts as the independent variables. The strongest predictors that were not multicollinear with other indices were related to local cohesion indices and syntactic complexity indices. These 12 indices showed medium to strong effect sizes (see Table 1 for the MANOVA results). These indices were used in the subsequent DFA.

Discriminant Function Analysis

A stepwise DFA using the indices selected through the MANOVA retained two variables from the 12 variables selected through the MANOVA as predictors. The remaining *ReaderBench* variables were removed. The two retained variables were related to local cohesion and syntactic complexity (see Table 2 for indices and unstandardized coefficients).

The results demonstrate that the DFA using these two indices correctly allocated 30 of the 38 texts in the total set, $\chi 2(df = 1) = 12.880$, $p < .001$, for an accuracy of 78.9%. The leave-one-out cross-validation (LOOCV) results reported the same classification accuracy (see the confusion matrix reported in Table 3 for results). The Cohen's Kappa

measure of agreement between the predicted and actual completion rate was .79, demonstrating substantial agreement.

Table 1. Study 1 – MANOVA results

Index	*F*	*η2*
Avg. simple subordinating conjunctions per sentence	17.561**	.328
Avg. nominal subject syntactic dependencies per sentence	15.572**	.302
Avg. indefinite pronouns per sentence	9.201*	.204
Avg. marker syntactic dependencies per paragraph	9.002*	.200
Avg. verbs per sentence	7.788*	.178
Avg. nominal subject syntactic dependencies per paragraph	6.692*	.157
Avg. adverbial clause modifier syntactic dependencies per sentence	6.128*	.145
Avg. sentence voice mutual information (dialogism)	5.746*	.138
Avg. clausal complement syntactic dependencies per paragraph	5.176*	.126
Avg. sentence relevance score	5.063*	.123
Avg. clausal complement syntactic dependencies per sentence	4.937*	.121
Avg. lexical chains per paragraph	4.782*	.117

* $p < .050$, ** $p < .001$.

Table 2. Study 1 – Discriminant Function Coefficients

Index	Coefficient
Avg. simple subordinating conjunctions per sentence	4.972
Avg. nominal subject syntactic dependencies per sentence	2.579

Constant = -5.578.

Table 3. Study 1 – Confusion matrix for DFA classifying texts

		Predicted	
	Actual	Original	Modified
Whole set	Original	16	3
	Modified	5	14
LOOCV	Original	16	3
	Modified	5	14

Discussion

The results of the DFA indicated that two *ReaderBench* indices were able to strongly distinguish between original texts and texts modified to increase cohesion. The results of the study demonstrate that the latter texts include more subordinating conjunctions and nominal subject dependencies. The first index is related to local cohesion (i.e., cohesion that connects smaller text segments such as sentences and phrases) and replicates findings in the original study which found that local cohesion in terms of lexical and semantic

overlap between sentences was a strong predictor of texts with increased cohesion (McNamara, Louwerse, McCarthy, & Graesser, 2010). The results also replicated a second study that used this dataset (Crossley, Kyle, & McNamara, 2016) which reported that texts modified to increase cohesion have an increased number of simple subordinators, reason and purpose connectives, and causal connectives. The second index that was predictive in this study (nominal subject dependencies) provides new information about increased text cohesion that was not evaluated in either McNamara, Louwerse, McCarthy, and Graesser (2010), or Crossley, Kyle, and McNamara (2016). Similar to McNamara et al., this finding indicates that the process of increasing text cohesion may have unintended effects on syntactic complexity. While relating to the *ReaderBench* indices, increasing connections between sentences and phrases (and inherently text cohesion) leads to an increase in the number of syntactic dependencies related to the head nouns (i.e., the number of phrases that the head noun controls). As an example, if two sentences such as (1) and (2):

1) The woman went to the store.
2) She needed milk.

are combined using a simple subordinator such as *because*, this will result in the sentence "*The woman went to the store because she needed milk*." While the cohesion in this sentence may increase, it seems to come at the cost of increased syntactic complexity. Nonetheless, added cohesion comes with substantial comprehension benefits for all ages, providing some evidence against theories that focus on processing resources (e.g., working memory), in contrast to theories of discourse comprehension (McNamara et al., 2010, 2014).

In terms of machine learning accuracy, the indices reported by *ReaderBench* led to a slightly higher classification accuracy than those reported by *Coh-Metrix* in the original study (McNamara, Louwerse, McCarthy, & Graesser, 2010) and the *Tool for the Automatic Analysis of Cohesion* (*TAACO*), which was used in the Crossley, Kyle, and McNamara (2016) study. For instance, the accuracy reported in McNamara at el. was 76% while the *ReaderBench* analysis was 79%. Lastly, the *Coh-Metrix* model reported in McNamara et al. reported the results for only the training set, which may have increased the reported accuracy. The accuracy found in Crossley et al. was lower than both the current *ReaderBench* analysis and Coh-Metrix model at 68%; however, it should be noted that *Coh-Metrix* relied on cohesion and word frequency, Crossley et al. only used cohesion indices, while the *ReaderBench* model included lexical, syntactic, semantic, and cohesion-centered indices.

Study 2 – Pairwise Text Comprehension Comparisons

Corpus

This study uses 150 news articles from the Guardian Weekly publication which were first reported by Crossley, Skalicky, Dascalu, Kyle, and McNamara (2017). The corpus included 50 original news articles, 50 that were simplified to the beginning level, and 50 simplified to the intermediate level. The articles were truncated using original paragraph breaks in order to ensure an approximately length of 150 words. Afterwards, these texts were ranked in terms of readability using the Mechanical Turk crowdsourcing service available through Amazon.com. Overall, 307 participants were recruited, and each evaluator provided approximately 10 pairwise comparisons of two texts with regards to text comprehension (which text was easier to understand), text processing (which text was quicker to read), and text familiarity (which addressed more familiar or knowledgeable topics). Crossley, Skalicky, Dascalu, Kyle, and McNamara (2017) applied a Bradley-Terry model (R. A. Bradley & Terry, 1952) to the 3,011 pairwise comparisons, for each of the previous three ranking criteria, to generate an evaluation of the text's difficulty in terms of its likelihood to be more difficult than all other texts. The Bradley-Terry model provides a global ranking of entries based on multiple individual pairwise comparisons (which in some cases can reflect opposite considerations of the raters); after maximizing the likelihood of the observed data and reaching convergence, texts that take longer to process are scored higher.

Statistical Analysis

Our interest in this *ReaderBench* analysis is to replicate the findings for the readability scores (i.e., the processing scores) reported by Crossley et al. (2017). To this end, we developed a regression model to predict text processing ratings derived from the Bradley-Terry models. Prior to the analysis, we removed any variables that violated a normal distribution to better assure that residuals in the regression model were distributed normally. Pearson correlations were then conducted on the remaining variables to determine whether they were meaningfully correlated with judgments of text processing. Any variables that did not report at least a small effect size (i.e., $r \geq .100$) with the text processing scores were removed from the analysis. The remaining variables were checked for multicollinearity to ensure that the final model consisted only of unique indices. For each pair of variables with absolute correlation values of $r >= .699$, only the variable with the highest correlation with text processing scores was retained. These variables were entered into a stepwise regression (bidirectional) using the stats package in R (R Core Team, 2013). Results were checked for significance multicollinearity using variance inflation values (VIF). The final model was checked for normality of residual, homoscedasticity, and constant error variance to ensure assumptions of linearity were met. The final model was then tested using 10-fold cross-validation through the caret

package (Kuhn, 2008) available in R. The method for inclusion into the model was also stepwise. Values reported for the model were co-efficient strengths and direction, relative importance metrics (predictors explained variance as non-negative contributions) using the relaimpo package (Grömping, 2006), and *t* and *p* values. The cross-validated model was also checked for all assumptions of linearity.

Results

We calculated correlations between the selected linguistic indices and the text processing ratings generated by the Bradley-Terry model. After controlling for normal distribution, effect sizes, multicollinearity, and multiple comparisons, the text processing judgment analysis included 51 linguistic indices. Correlations are presented in Table 4.

To analyze which linguistic features predicted the text processing ratings, we conducted a stepwise regression analysis using the selected linguistic indices as the independent variables. This yielded a significant model, $F(9, 140) = 10.440$, $p < .001$, $R^2 = .402$. Nine variables were significant predictors of the text comprehension ratings. These nine variables, when used in a 10-fold cross-validation model yielded a model with an $R^2 = .301$ and RMSE = .005 (see Table 5 for coefficients, *t* values, and *p* values for the model).

DISCUSSION

A unique feature of this analysis and the database is that it focused on text processing and not text comprehension (i.e., how well a text is understood). Most readability studies focus on comprehension (e.g., Kate et al., 2010) and not processing even though most theories of text readability include text processing within their definition (Chall & Dale, 1995; Newbold & Gillam, 2010). Thus, this study helps to build on the few studies that focus on text processing as a function of text readability (cf. Crossley et al., 2017).

The results of the regression model indicate that *ReaderBench* variables related to sentence length, cohesion, pronoun use, dialogism, and syntactic dependencies were all related to text processing speed. The coefficients suggest that news articles that took longer to read had longer sentences, had a greater number of referents per paragraph, had great paragraph cohesion, and were more syntactically complex. Texts were also more quickly read if they had greater paragraph to text cohesion. These findings support models of text readability which postulate that syntactic and discourse features are important components of readability (Just & Carpenter, 1980; Koda, 2005) and that greater complexity in these features will lead to greater difficulty in text processing.

Table 4. Study 2 – Correlations with reading processing scores

Index	*r*	*p*
Avg. nouns per sentence	.429	<.001
Average sentence length expressed in characters	.411	<.001
Avg. case syntactic dependencies per sentence	.357	<.001
Avg. compound syntactic dependencies per sentence	.356	<.001
Avg. verbs per sentence	-.328	<.001
Avg. rhythmic units per sentence	.297	<.001
Avg. sentence-paragraph cohesion using Wu-Palmer semantic distance in WordNet	.289	<.001
Avg. distance between lemma and word stems (characters)	.282	<.001
Avg. punctuation syntactic dependencies per sentence	.278	<.001
Avg. distance between words and corresponding stems (characters)	.273	<.001
Avg. word AoA scores (Schock, Cortese, Khanna, & Toppi, 2012) per paragraph	-.264	<.001
Avg. sentences per paragraph	.263	<.001
Avg. copula syntactic dependencies per paragraph	-.255	<.010
Avg. adjectives per sentence	.238	<.010
Avg. word AoE scores using index above threshold (0.3) per paragraph	.236	<.010
Avg. unique pronouns per paragraph	.231	<.010
Avg. word AoA scores (Bird, Franklin, & Howard, 2001) per sentence	-.226	<.010
Avg. sentence-paragraph cohesion using cosine similarity in LSA vector spaces	.225	<.010
Avg. sentence-paragraph cohesion using the inverse Jensen-Shannon dissimilarity between LDA probability distributions	.214	<.010
Avg. coordination syntactic dependencies per paragraph	.213	<.010
Avg. open clausal complement syntactic dependencies per sentence	-.212	<.010
Avg. commas per sentence	.210	<.010
Avg. coordinating conjunctions per paragraph	.185	<.050
Avg. word AoE scores using the inverse linear regression slope per paragraph	.178	<.050
Avg. word AoA scores (Bird, Franklin, & Howard, 2001) per paragraph	-.174	<.050
Avg. logical connectors per paragraph	.168	<.050
Avg. co-references per chain	.164	<.050
Avg. distribution of voices per sentence (dialogism)	-.163	<.050
Avg. voice entropy per paragraph (dialogism)	.163	<.050
Avg. sentence linking connectors per paragraph	.159	<.050
Avg. word AoE scores using the inverse linear regression slope per sentence	-.145	<.050
Avg. document flow cohesion using cosine similarity in LSA vector spaces and maximum criterion	.143	<.050
Avg. paragraph-document cohesion using cosine similarity in LSA vector spaces	-.141	<.050
Avg. document flow cohesion using the Jensen-Shannon dissimilarity between LDA probability distributions and maximum criterion	-.141	<.050
Maximum flow ordered sequence using cosine similarity in word2vec spaces and above mean plus standard deviation criterion	-.140	<.050
Avg. word polysemy count (only content words)	.137	<.050
Maximum flow ordered sequence using the Wu-Palmer semantic distance in WordNet and above mean plus standard deviation criterion	-.137	<.050
Maximum flow ordered sequence using the Jensen-Shannon dissimilarity between LDA probability distributions and above mean plus standard deviation criterion	-.131	<.050
Standard deviation of sentence relevance scores	-.129	<.050
Spearman correlation of flow versus initial ordering using the cosine similarity in LSA	.129	<.050

Index	*r*	*p*
vector spaces and maximum criterion		
Maximum flow ordered sequence using the Jensen-Shannon dissimilarity between LDA probability distributions and maximum criterion	.126	<.050
Avg. simple subordinating conjunctions per paragraph	-.120	<.050
Avg. stressed syllables in rhythmic unit	-.118	<.050
Avg. sentence-paragraph cohesion using cosine similarity in word2vec spaces	.117	<.050
Avg. syllables in a rhythmic unit	.117	<.050
Absolute position accuracy based on topological sort using cosine similarity in LSA vector spaces and above mean plus standard deviation criterion	-.113	<.050
Maximum voice span (dialogism)	.108	<.050
Avg. paragraph adjacency cohesion using the Jensen-Shannon dissimilarity between LDA probability distributions	.103	<.050

Table 5. Study 2 – Summary of multiple regression model for pairwise comparisons (text processing)

	Entry	*Relative importance*	Coefficient	*t*	*p*
1	Average sentence length expressed in characters	0.2455	0.001	3.110	.002
2	Avg. unique pronouns per paragraph	0.1594	0.011	4.260	<.001
3	Avg. case syntactic dependencies per sentence	0.1591	0.001	1.790	.076
4	Avg. punctuation syntactic dependencies per sentence	0.1014	0.004	2.000	.047
5	Average paragraph adjacency cohesion using the Jensen-Shannon dissimilarity between LDA probability distributions	0.0746	0.001	3.150	.002
6	Avg. paragraph-document cohesion using cosine similarity in LSA vector spaces	0.0706	-0.003	-3.040	.003
7	Avg. co-references per chain	0.0687	0.001	2.640	.009
8	Avg. voice entropy per paragraph (dialogism)	0.0656	0.001	2.500	.013
9	Maximum voice span (dialogism)	0.0552	0.001	2.390	.018

Note: Constant = -0.062.

In comparison to Crossley et al. (2017), *ReaderBench* indices alone accounted for less variance (i.e., 30% as compared to 47% of the variance) but, again, differences between the two studies can explain these disparities. For instance, Crossley et al. used NLP tools that reported on phrasal indices, one of which (tri-gram incidence) explained 20% of the text processing variance in the original study. Such measures are not currently included in *ReaderBench*.

The *ReaderBench* indices in this study also make unique contributions to our understanding of text processing. For instance, to our knowledge, the number of unique pronouns per paragraph, coupled with the number of co-references per chain, has not been a significant predictor of text processing in previous studies. This finding suggests that texts with a greater number of referents are more difficult to process. While unique, the finding does overlap with the model reported in Crossley et al. (2017) which reported that a greater number of entities per sentence (e.g., people's names, venues,

organizations) leads to slower processing. Moreover, indices derived from dialogism sustaining the diversity (i.e., entropy of different points of view) and spread (i.e., span) of ideas were also predictive, in tight connection with the indices related to global cohesion. Lastly, this *ReaderBench* analysis provides evidence that syntactic complexity (i.e., punctuation and case syntactic dependencies per sentence) in a text can decrease text processing speed. While theorized, little evidence has supported this assertion.

CONCLUSION

To our knowledge, *ReaderBench* is a unique open-source, multilingual framework, which provides a wide range of textual complexity indices and that can be used to perform various text cohesion and in-depth discourse analyses. It combines several semantic and discourse analysis techniques from Natural Language Processing into CNA and implements the ideas of the polyphonic model together with its associated ideas, such as rhythm and interanimation. The tool and the corresponding web services are available both online (http://readerbench.com) and for download as a desktop client that includes additional functionalities to the ones published online (http://readerbench.com/deployment).

The results from the first study indicate that the textual complexity indices from *ReaderBench* can differentiate texts among different levels of cohesion that plays an important role in comprehension. The second study provides evidence for *ReaderBench*'s wide applicability with respect to the high number of significantly correlated textual complexity indices, most of which are centered on semantics and discourse. These findings demonstrate *ReaderBench's* versatility and generalizability to multiple contexts (Botarleanu, Dascalu, Sirbu, Crossley, & Trausan-Matu, 2018). As such, learners and educators can potentially use the ReaderBench framework across multiple pedagogical scenarios.

ACKNOWLEDGMENTS

This research was partially supported by the 644187 EC H2020 *Realising an Applied Gaming Eco-system* (RAGE) project, the Department of Education, Institute of Education Sciences - Grant R305A130124, as well as by the Department of Defense, Office of Naval Research - Grants N00014140343 and N000141712300.

REFERENCES

Allen, L. K., Dascalu, M., McNamara, D. S., Crossley, S., & Trausan-Matu, S. (2016). Modeling Individual Differences among Writers Using ReaderBench. In *8th Int. Conf. on Education and New Learning Technologies (EduLearn16)* (pp. 5269–5279). Barcelona, Spain: IATED.

Balint, M., Dascalu, M., & Trausan-Matu, S. (2016a). Classifying Written Texts through Rhythmic Features. In *15th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2016)* (pp. 121–129). Varna, Bulgaria: Springer.

Balint, M., Dascalu, M., & Trausan-Matu, S. (2016b). The Rhetorical Nature of Rhythm. In *15th Int. Conf. on Networking in Education and Research (RoEduNet)* (pp. 48–53). Bucharest, Romania: IEEE.

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers, 33*(1), 73–79.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Botarleanu, R. M., Dascalu, M., Sirbu, M. D., Crossley, S. A., & Trausan-Matu, S. (2018). ReadME – Generating Personalized Feedback for Essay Writing using the ReaderBench Framework. In *3rd Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2018)* (pp. 133–145). Aalborg, Denmark: Springer.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings.* Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika, 39*, 324–345.

Cambria, E., Grassi, M., Poria, S., & Hussain, A. (2013). Sentic computing for social media analysis, representation, and retrieval. In Ramzan, N., Zwol, R., Lee, J. S., Clüver, K. & Hua, X. S. (Eds.), *Social Media Retrieval* (pp. 191–215). New York, NY: Springer.

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

Crossley, S. A., Dascalu, M., Trausan-Matu, S., Allen, L., & McNamara, D. S. (2016). Document Cohesion Flow: Striving towards Coherence. In *38th Annual Meeting of the Cognitive Science Society* (pp. 764–769). Philadelphia, PA: Cognitive Science Society.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227–1237.

Crossley, S. A., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining Click-Stream Data with NLP Tools to Better Understand MOOC Completion. In *6th Int. Conf. on Learning Analytics & Knowledge (LAK '16)* (pp. 6–14). Edingurgh, UK: ACM.

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In Biswas, G., Bull, S., Kay, J., & Mitrovic, A. (Eds.), *15th Int. Conf. on Artificial Intelligence in Education* (pp. 438–440). Christchurch, New Zealand: Springer.

Crossley, S. A., Skalicky, S. C., Dascalu, M., Kyle, K., & McNamara, D. S. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54*(5-6), 340–359.

Dascalu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating*. Cham, Switzerland: Springer.

Dascalu, M., Allen, K. A., McNamara, D. S., Trausan-Matu, S., & Crossley, S. A. (2017). Modeling Comprehension Processes via Automated Analyses of Dialogism. In *39th Annual Meeting of the Cognitive Science Society (CogSci 2017)* (pp. 1884–1889). London, UK: Cognitive Science Society.

Dascalu, M., Dessus, P., Bianco, M., & Trausan-Matu, S. (2014). Are Automatically Identified Reading Strategies Reliable Predictors of Comprehension? In S. Trausan-Matu, K. E. Boyer, M. Crosby & K. Panourgia (Eds.), *12th Int. Conf. on Intelligent Tutoring Systems (ITS 2014)* (pp. 456–465). Honolulu, USA: Springer.

Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A. (2014). Mining texts, learner productions and strategies with ReaderBench. In Peña-Ayala, A. (Ed.), *Educational Data Mining: Applications and Trends* (pp. 345–377). Cham, Switzerland: Springer.

Dascalu, M., Gifu, D., & Trausan-Matu, S. (2016). What Makes your Writing Style Unique? Significant Differences between Two Famous Romanian Orators. In Nguyen, N. T., Manolopoulos, Y., Iliadis, L., & Trawinski, B. (Eds.), *8th Int. Conf. on Computational Collective Intelligence (ICCCI 2016)* (pp. 143–152). Halkidiki, Greece: Springer.

Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., McNamara, D. S., Crossley, S., & Trausan-Matu, S. (2017). ReaderBench: A Multi-Lingual Framework for Analyzing Text Complexity. In Lavoué, E., Drachsler, H., Verbert, K., Broisin, J., & Pérez-Sanagustín, M. (Eds.), *12th European Conference on Technology Enhanced Learning (EC-TEL 2017)* (pp. 495–499). Tallinn, Estonia: Springer.

Dascalu, M., McNamara, D. S., Crossley, S. A., & Trausan-Matu, S. (2015). Age of Exposure: A Model of Word Learning. In *30th AAAI Conference on Artificial Intelligence* (pp. 2928–2934). Phoenix, AZ: AAAI Press.

Dascalu, M., McNamara, D. S., Trausan-Matu, S., & Allen, L.K. (2018). Cohesion Network Analysis of CSCL Participation. *Behavior Research Methods, 50*(2), 604–619. doi: 10.3758/s13428-017-0888-4

Dascalu, M., Popescu, E., Becheru, A., Crossley, S. A., & Trausan-Matu, S. (2016). Predicting Academic Performance Based on Students' Blog and Microblog Posts. In *11th European Conference on Technology Enhanced Learning (EC-TEL 2016)* (pp. 370–376). Lyon, France: Springer.

Dascalu, M., Stavarache, L. L., Trausan-Matu, S., Dessus, P., & Bianco, M. (2014). Reflecting Comprehension through French Textual Complexity Factors. In *26th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2014)* (pp. 615–619). Limassol, Cyprus: IEEE.

Dascalu, M., Trausan-Matu, S., McNamara, D. S., & Dessus, P. (2015). ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395–423. doi: 10.1007/s11412-015-9226-y.

Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., & Kurvers, H. (2017). ReaderBench Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch. In Baker, A. E. R., Hu, X., Rodrigo, M. M. T. & du Boulay, B., (Eds.), *18th Int. Conf. on Artificial Intelligence in Education (AIED 2017)* (pp. 52–63). Wuhan, China: Springer.

Gervasi, V., & Ambriola, V. (2002). Quantitative assessment of textual complexity. In M. L. Barbaresi (Ed.), *Complexity in language and text* (pp. 197–228). Pisa, Italy: Plus.

Gifu, D., Dascalu, M., Trausan-Matu, S., & Allen, L. K. (2016). Time Evolution of Writing Styles in Romanian Language. In *28th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2016)* (pp. 1048–1054). San Jose, CA: IEEE.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers, 36*(2), 193–202.

Grömping, Ulrike. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software, 17*(1), 1–27.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics, 21*(2), 203–225.

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

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., & Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *23rd Int. Conf. on Computational Linguistics* (pp. 546–554): Association for Computational Linguistics.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge, MA: Cambridge University Press.

Kuhn, M. (2008). Caret package. *Journal of statistical software, 28*(5), 1–26.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990.

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *Modern Language Journal, 102* (2), 333-349.

Kyle, K., Crossley, S. A., & Berger, C. (in press). The Tool for the Automatic Analysis of Lexical Sophistication Version 2.0. *Behavior Research Methods*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*(2), 211–240.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1), 92–108.

Lasswell, H. D., & Namenwirth, J. Z. (1969). *The Lasswell Value Dictionary*. New Haven: Yale University Press.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., & Jurafsky, D. (2011). Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. In *Fifteenth Conference on Computational Natural Language Learning: Shared (TaskCONLL Shared Task '11)* (pp. 28–34). Portland, OR: ACL.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). Baltimore, MA: ACL.

Marcus, S. (1970). *Poetica matematică*. Bucharest, Romania: Editura Acad. Rep. Soc. Romania.

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

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across the ages and genres. In Sabatini, J. P., Albro, E. & O'Reilly, T. (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Lanham, MD: R&L Education.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representation in Vector Space. In *Workshop at ICLR*. Scottsdale, AZ.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.

Newbold, N., & Gillam, L. (2010). The linguistics of readability: the next step for word processing. In *NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids* (pp. 65–72). Los Angeles, CA: Association for Computational Linguistics.

Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.

Page, E. (1968). Analyzing student essays by computer. *International Review of Education, 14*(2), 210–225.

Pennebaker, James W, Booth, Roger J, & Francis, Martha E. (2007). *Linguistic inquiry and word count: LIWC* [Computer software]. Austin, TX: University of Texas.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping e-rater®: Challenging the validity of automated essay scoring*. Princeton, NJ: Educational Testing Service.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: Foundation for Statistical Computing.

Roscoe, R. D., Varner, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social science information, 44*(4), 695–729.

Schock, J., Cortese, M. J., Khanna, M. M., & Toppi, S. (2012). Age of acquisition estimates for 3,000 disyllabic words. *Behavior Research Methods, 44*(4), 971–977.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, 27*, 379–423 & 623–656.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal, 30*, 50–64.

Stone, P., Dunphy, D. C., Smith, M. S., Ogilvie, D. M., & associates. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge, MA: The MIT Press.

Trausan-Matu, S., Dascalu, M., & Rebedea, T. (2014). PolyCAFe–automatic support for the polyphonic analysis of CSCL chats. *International Journal of Computer-Supported Collaborative Learning, 9*(2), 127–156. doi: 10.1007/s11412-014-9190-y.

Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2007). Supporting polyphonic collaborative learning. *E-service Journal, 6*(1), 58–74.

Wresch, W. (1993). The imminence of grading essays by computer—25 years later. *Computers and Composition, 10*(2), 45–58.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics, ACL '94* (pp. 133–138). New Mexico, USA: ACL.