Journal of Research in Reading
UKLA
The United Kingdom Literacy Association

Journal of Research in Reading, ISSN 0141-0423
Volume 42, Issue 3-4, 2019, pp 541–561
DOI:10.1111/1467-9817.12283

Moving beyond classic readability formulas: new methods and new models

Scott A. Crossley

Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA, USA

Stephen Skalicky

School of Linguistics and Applied Language Studies, Victoria University of Wellington, Wellington, New Zealand

Mihai Dascalu

Department of Computer Sciences, Politehnica University of Bucharest, Bucharest, Romania

Background: Advances in natural language processing (NLP) and computational linguistics have facilitated major improvements on traditional readability formulas that aim at predicting the overall difficulty of a text. Recent studies have identified several types of linguistic features that are theoretically motivated and predictive of human judgments of text readability, which outperform predictions made by traditional readability formulas, such as Flesch–Kincaid. The purpose of this study is to develop new readability models using advanced NLP tools to measure both text comprehension and reading speed.
Methods: This study used crowdsourcing techniques to collect human judgments of text comprehension and reading speed across a diverse variety of topic domains (science, technology and history). Linguistic features taken from state-of-the-art NLP tools were used to develop models explaining human judgments of text comprehension and reading speed. The accuracy of these models was then compared with classic readability formulas.
Results: The results indicated that models employing linguistic features more theoretically related to text comprehension and reading speed outperform classic readability models.
Conclusions: This study developed new readability formulas based on advanced NLP tools for both text comprehension and reading speed. These formulas, based on linguistic features that better represent theoretical and behavioural accounts of the reading process, significantly outperformed classic readability formulas.

Keywords: readability, natural language processing, crowdsourcing, text comprehension, text reading speed

Highlights

What is already known about this topic

- Classic readability formulas are widely used to assess the difficulty of texts.
- Classic readability formulas lack construct validity.
- Classic readability formulas are predictive on the data sets on which they have been trained but do not perform as well on new data sets.

What this paper adds

- This study introduces NLP methods to better link text features with text complexity constructs that have strong overlap with theories of reading.
- This study introduces techniques to collect readability benchmarks (crowdsourcing techniques).
- This paper introduces new readability formulas to assess text comprehension and reading speed.

Implications for theory, policy or practice

- The formulas developed provide evidence linking text complexity to theories of text comprehension and reading speed.
- The formulas are stronger predictors of text comprehension and reading speed than traditional readability formulas and can be scaled.
- The paper provides new evidence that classic readability formulas are not strong predictors of either text comprehension or reading speed.

Developing proficient reading skills is an important component of success, not only in academic settings but also in business and social settings (Geiser & Studley, 2002; Powell, 2009). However, reading skills are often difficult to acquire (National Assessment of Educational Progress, 2011), and many students perform below average on standardised reading tests at primary, secondary and tertiary levels (U. S. Department of Education, 2003). One approach to improve reading skills is to match readers with texts that challenge them but are not incomprehensible or too difficult to read. When readers are matched to texts of sufficient difficulty, reading skills can improve (Allington, 2005; Stanovich, 1985; Wolfe et al., 1998).

A common method to match readers to appropriate texts is the use of readability formulas. Such formulas are manifold, with well over 200 models developed since the 1940s (Benjamin, 2012). Generally, these formulas rely on superficial text-based features to assess readability including the number of words per sentence, which is meant to act as a proxy for syntactic complexity, and the number of characters per word, which is meant to act as a proxy for lexical difficulty. While both syntactic complexity and lexical difficulty are important components of readability (Just & Carpenter, 1980), sentence and word length measures likely do not tap directly into linguistic components related to readability (Crossley, Greenfield, & McNamara, 2008) nor are they the only linguistic features related to readability. For instance, text cohesion is an important component of readability (Gernsbacher, 1990), but it is not measured in the majority of readability formulas. For

these reasons, many readability formulas are argued to have weak construct validity (Davison & Kantor, 1982). They do, however, report strong correlations with text comprehension criteria in a number of studies (Chall & Dale, 1995; Fry, 1989).

Even though many readability formulas have a variety of weaknesses, they have been adopted by a number of institutions including primary and secondary schools, universities, the military and testing agencies (to name but a few) where they are used to select reading material presumed to be appropriate for a variety of reading skills (DuBay, 2004). Unfortunately, in many cases, readability formulas are used to select texts for readers and reading situations that go well beyond the limitations of the formulas, as evidenced by a number of recent studies indicating that classic readability formulas (i.e., Flesch Reading Ease, Flesch–Kincaid Grade Level and Dale–Chall) perform worse than other models for different reading tasks, reading domains and populations (Begeny & Greene, 2014; Crossley et al., 2008; Crossley et al., 2017; François & Fairon, 2012; François & Miltsakaki, 2012; Hiebert, 2011; Kate et al., 2010; Newbold & Gillam, 2010; Pitler & Nenkova, 2008). For instance, recent studies by Crossley et al. (2017) and De Clercq, Hoste, Desmet, Van Oosten, De Cock, and Macken (2014) indicated that classic readability formulas were less predictive of comprehension than readability formulas developed using linguistic features based on word phrases (in the De Clercq et al. study) and on features that measure lexical and syntactic constructs, text cohesion, sentiment, topic analysis and semantics (in the Crossley et al. study). Additionally, unlike many previous studies, Crossley et al. also examined reading speed, which reported low correlations with classic readability formulas and significantly lower values in contrast to newly developed formulas. However, as mentioned by Crossley et al., additional studies that examine larger reading data sets are needed.

Thus, this study builds on the Crossley et al. (2017) and De Clercq (2014) studies by developing new text comprehension and reading speed formulas based on larger reading data sets selected from multiple topic domains using more sophisticated linguistic features. The goal of this study is to develop text complexity formulas that better match cognitive models of reading by taking into consideration discourse-specific features, such as situation level and text-based processing (Kintsch, Welsch, Schmalhofer, & Zimny, 1990) that better explain how text features interact with readers (Kintsch, 1994; Miller & Kintsch, 1980).

Text readability

Text readability is best defined as the ease with which a text can be read and understood in terms of the linguistic features found within a text (Dale & Chall, 1949; Richards et al., 1992). However, in practice, most readability studies and formulas focus primarily on readability as a measure of text understanding (e.g., Kate et al., 2010) and not text processing (i.e., reading speed), although many definitions of text readability incorporate ease of text processing as an important component. For example, Newbold and Gillam (2010) defined readability as a measure of how different readers could ‘comfortably read a text’ (p. 66), whereas Richards et al. (1992) referred to it as ease of reading and understanding texts. One of the first definitions to include text processing as a measure of readability was likely Dale and Chall (1949), who discussed readability in terms that included not only text understanding but also reading speed. However, most classic readability formulas are derived from comprehension data (i.e., text understanding) and not text processing/reading speed data.

Theoretical accounts of text comprehension and processing generally associate three linguistic features with readability: lexical sophistication, syntactic complexity and discourse structures (Just & Carpenter, 1980; Snow, 2002). These features are related to text elements inherent in text that contribute to text complexity. For instance, numerous studies have indicated that lexical sophistication features such as sound and spelling relationships between words (Juel & Solso, 1981; Mesmer, 2005), word familiarity and frequency (Howes & Solomon, 1951), and word imageability and concreteness (Richardson, 1975) lead to faster processing and more accurate word decoding. Beyond lexical sophistication features, the meaning of words (i.e., semanticity) is also an important component of text readability because readers need to be able to process and decode words, as well as know the meaning of the decoded words (Mesmer, Cunningham, & Hiebert, 2012). Thus, the semanticity of words and larger text segments can help readers link common themes or ideas as well as process and understand words more easily based on background knowledge and text familiarity (Bailin & Grafstein, 2001; McNamara & Kintsch, 1996). Semantic features that can be derived from the text include topics or named entities (Newbold & Gillam, 2010; Zamanian & Heydari, 2012).

As words are read, effective readers are able to parse these words into syntactic structures that help organise main ideas and assign thematic roles to arguments (Graesser, Swamer, Baggett, & Sell, 1996; Just & Carpenter, 1980; Mesmer et al., 2012). In terms of text complexity, a number of text features allow for quicker syntactic parsing including words or morphemes per t-unit[1] (Cunningham et al., 2005) or sentence length (Klare, 1984). As syntactic representations are being developed, readers continuously form larger larger discourse structures to create a discourse thread (Grimes, 1975). These discourse structures are related to text cohesion and can be partially constructed using linguistic features in the text that link words and concepts within and across syntactic structures (Givón, 1995). Sensitivity to cohesion structures help readers process and understand paragraphs and larger discourse segments (Gernsbacher, 1990; Mesmer et al., 2012) because they allow readers to develop and build relationships between text elements (e.g., words, sentences or paragraphs), which can help construct knowledge representations (Britton & Gülgöz, 1991; Kintsch, 1988; McNamara & Kintsch, 1996).

In terms of readability formulas, many classic formulas incorporate some aspect of word sophistication and syntactic complexity (Chall & Dale, 1995; Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975). However, these formulas often only use proxy measures (i.e., number of characters per word and number of words per sentence)[2] and they generally ignore semantic features and discourse structures. Moreover, most classic readability formulas are based on text comprehension scores typically collected using cloze tests or multiple-choice questions, both of which are unlikely to measure deep comprehension (Magliano, Millis, Ozuru, & McNamara, 2007). Solely developing formulas that are predictive of comprehension alone also ignore text reading speed, which is strongly related to text readability (Bailin & Grafstein, 2001; Crossley et al., 2017; Dale & Chall, 1949; Newbold & Gillam, 2010; Richards et al., 1992).

Natural language processing tools and readability formulas

More recent readability work has begun to harness the power of advanced natural language processing (NLP) tools to derive more principled and conceptually valid language features related to text complexity that can be used to examine text difficulty (Benjamin, 2012; Crossley, Dufty, McCarthy, & McNamara, 2007; Crossley et al., 2008, Crossley et al.,

2017; Graesser, McNamara, & Kulikowich, 2011). These newer NLP tools report on a variety of indices related to lexical sophistication, semanticity, syntactic complexity and text cohesion that have much stronger overlap with text complexity features that are important for both comprehension and reading speed.

For instance, Schwarm and Ostendorf (2005) used a parser to calculate the number of verb and noun phrases, parsing tree depth and embedded clauses in a text and used these features to successfully develop a readability formula that discriminated between text reading levels. Collins-Thompson and Callan (2005) used a text-classification approach based on the probabilities of words appearing in different grade levels (as opposed to basic word frequency) to accurately classify different web texts by grade levels. Feng, Jansche, Huenerfauth, and Elhadad (2010) used similar text-classification methods and found that combining fine-grained NLP measures (e.g., lexical chains and part-of-speech information), together with features included in traditional readability formulas, significantly increased text-classification accuracy. Heilman, Collins-Thompson, Callan, and Eskenazi (2006) examined links between the frequency of grammatical constructions and grade level and found that models without grammatical features had higher classification errors. Crossley et al. (2007, 2008) used measures of syntactic complexity, word frequency and text cohesion to predict text readability for L1 (2007) and L2 (2008) readers. These studies reported that the developed L1 readability formulas performed as well as classic readability formulas and that the L2 formulas outperformed classic readability formulas. Pitler and Nenkova (2008) also combined syntactic, lexical and discourse features to predict judgments of readability, reporting that these features were strong predictors whereas the features found in classic formulas were not. Lastly, Newbold and Gillam (2010) proposed readability metrics based on linguistic indices related to textual cohesion, propositional density and word frequency. They reported correlations only between the word frequency metrics and traditional readability formulas and found that the word frequency measure correlated highly with traditional readability measures.

More recently, researchers have begun to apply NLP techniques on corpora collected using crowdsourcing methods to develop readability formulas. For instance, De Clercq et al. (2014) asked experts (e.g., professors) and nonexperts (e.g., online workers) to provide pairwise judgments of text difficulty using a large corpus of generic Dutch texts and then developed a readability model based on the probability of one text being ranked as easier or more difficult than another text. De Clercq et al. reported high, positive correlations between expert and nonexpert ratings, as well as between the crowdsourced ratings and traditional readability formulas. They also reported differences in text readability as a function of domain (e.g., administrative texts, manuals and news articles). Lastly, they found that readability formulas derived from n-gram features reported significantly higher accuracy in predicting the pairwise human ratings of texts than traditional readability ratings. More recently, Crossley et al. (2017) used NLP tools to examine text readability for a corpus of 150 news articles. This study relied on a pairwise comparison algorithm to rank texts based on crowdsourced human judgments of difficulty and asked raters to judge which of the two texts was easier to understand (i.e., text comprehension) and which of the two texts they read more quickly (i.e., text reading speed). The crowdsourcing approach allowed Crossley et al. to develop reading criteria on a much larger data set than previous studies (with the corpus size double to quadruple that used to develop most readability formulas). The models derived from the NLP tools for both comprehension and reading speed significantly outperformed classic readability formulas. Not only that, the NLP indices that were predictive of reading criteria supported theoretical and

psycholinguistic models of reading. For instance, the comprehension model indicated that more easily understood texts contained less sophisticated words, lower text cohesion and fewer verbs. The text reading speed model indicated that texts that were easier to read contained fewer unique trigrams (i.e., three-word phrases), less sophisticated words and fewer entities (e.g., proper nouns) per sentence.

Current study

The current study builds on the work by De Clercq et al. (2014) and Crossley et al. (2017) by developing models that predict text comprehension and reading speed using state-of-the-art NLP tools. Our approach also uses crowdsourcing techniques to collect human judgments of text comprehension and reading speed across a diverse variety of topic domains (science, technology and history). However, the sample size in this study is significantly larger than that reported by De Clercq et al. and Crossley et al. Like Crossley et al., pairwise comparisons among the ratings were calculated using a Bradley–Terry model to estimate the difficulty of each text in comparison with all the other texts. We used a number of NLP features to develop models for text comprehension and reading speed. The accuracy of these models was then compared with classic readability formulas.

The research questions that guide this study are the following:

1 Which linguistic features are predictive of pairwise scores for text comprehension and reading speed obtained through crowdsourcing methods?
2 Are models that incorporate these linguistic features more predictive of text comprehension and reading speed than classic text readability formulas?

We selected classic text readability formulas (i.e., Flesch Reading Ease, Flesch–Kincaid Grade Level and the New Dale–Chall Readability Formula) as our baseline models over newer approaches to assess text complexity because the majority of contemporary research does not provide standardised readability formulas that can be replicated (e.g., Collins-Thompson & Callan, 2005; Feng et al., 2010; Heilman et al., 2006; Newbold & Gillam, 2010; Pitler & Nenkova, 2008; Schwarm & Ostendorf, 2005). Crossley et al. (2008) did provide a standardised readability formula, but it was developed for second language readers, which is inappropriate for this study. Beyond not providing standardised readability formulas, many of the studies predict grade level as a proxy for readability (Collins-Thompson & Callan, 2005; Collins-Thompson et al., 2006; Feng et al., 2010; Schwarm & Ostendorf, 2005). In addition, most do not provide enough information on how the NLP features were calculated to allow for replication (Newbold & Gillam, 2010; Schwarm & Ostendorf, 2005), rely on NLP tools that are no longer supported (Feng et al., 2010) or use bag-of-word methods that are specific to the analysed corpus (Collins-Thompson & Callan, 2005). In contrast, classic readability formulas are readily available and are easy to replicate.

Method

Corpus

The corpus for this project composed of the introduction sections of articles from the simplified and authentic English versions of the online encyclopaedia *Wikipedia*. While the

authentic version of *Wikipedia* allows authors linguistic autonomy, instructions for authors on the simplified version of *Wikipedia* advise them to 'use basic English vocabulary and shorter sentences' in order to help readers 'understand normally complex terms or phrases' (Wikipedia, n.d.). We selected both simplified and authentic entries to include texts of different difficulty levels, from the same domain. A total of 600 texts were selected, each of approximately 200 words long. The corpus was split evenly between the simplified and authentic versions of *Wikipedia* (n = 300). We then evenly selected texts from three different topic domains (i.e., science, technology and history) from each version of *Wikipedia*. Each of the 600 articles was about a unique topic.

Participants

We recruited online participants through Amazon Mechanical Turk (AMT), a crowdsourcing website where workers are paid small fees in exchange for anonymous, online work. One of the benefits of using AMT is the ability to recruit a large number of participants in a relatively short amount of time, allowing us to collect a large number of pairwise comparisons for our *Wikipedia* texts. Studies of the AMT worker population have demonstrated that it is more diverse than traditional university undergraduate research pools and that, with basic quality control mechanisms in place, the data yielding from AMT studies are just as reliable (Buhrmester, Kwang, & Gosling, 2011; Goodman, Cryder, & Cheema, 2013). Accordingly, we used TurkPrime, a third party website that ensures quality of AMT data by confirming IP addresses and blocking duplicate workers (Litman, Robinson, & Abberbock, 2016). In addition, previous studies of readability have shown that crowdsourced readability ratings from nonexperts correlated signficantly with those of expert raters (r = .86–.90; De Clercq et al., 2014), suggesting that layperson ratings are as valid as expert ratings.

For this study, we recruited 915 participants from the United States and obtained 9,194 text judgments from them (each participant was recruited to complete at least 10 judgments; some workers completed more, while some completed fewer because of errors with web browsers and other technological issues). Each worker was paid $1.50 for their participation. Sixty participants were removed for not answering at least 75% of the recall questions correctly (see succeeding text). Thus, in total, 855 participants provided 8,605 pairwise comparisons that were used to derive text comprehension and reading speed ratings for each text.

Study design

We designed a custom website to present the pairs of texts to the participants and collect the data. The website first collected data from participants pertaining to their demographic information (gender, age and first language), as well as reading and television habits (number of books read per year, hours of television watched per day, how much they enjoy reading on a scale of 1–5 and confidence in reading ability on a scale of 1–6). Seventeen participants indicated that their first language was not English and reported first languages including Dutch, Esan, French, Indonesian, Mandarin Chinese, Russian, Spanish, Tagalog, Telugu, Turkish, Ukrainian and Vietnamese. Reported genders for the participants included 462 women and 393 men, with an average age of 37.65 years. On average,

Table 1. Demographic features of Amazon Mechanical Turk participants (n = 855).

	M	SD
Books read per year	4.40	1.92
Hours of television per day	3.06	1.86
Age	37.65	11.61
Enjoy reading	4.22	0.91
Confidence in reading	5.36	1.02

Note. Participants answered how much they enjoyed reading using a scale of 1–5, with 1 being *not at all* and five being *very much*. The confidence in reading scale ranged from 1 to 6, with a rating of 1 meaning *very unconfident* and a score of six meaning *very confident*.

participants reported reading 4.4 books per year and watched 3 hours of television per day. See Table 1 for detailed survey results.

After completing the survey, the participants made 10 text comparisons selected from among the 600 text corpus. For each comparison, two randomly selected texts were presented side by side on the screen. Instructions underneath the two texts informed participants to first read the texts and then answer a series of questions about the text. The first two questions were true/false recall questions, one per text. The text recall questions were all in the form of 'The text mentions ____' and were used as accuracy criteria for inclusion and exclusion of participants in the final data set. The recall questions were simple and were not meant to assess comprehension. Rather, they were used to assess whether the participants read the text in order to filter out participants who did not take the task seriously. Participants were also asked to choose which of the two texts was easier to understand (i.e., comprehension), which of the two texts they read more quickly (i.e., how much processing effort participants felt they had put into reading the text) and which of the two texts they were more familiar with (see Figure 1 for overview of survey). The first two comparisons formed the data used in subsequent readability analyses. Participants repeated this process for 10 sets of two texts, with none of the texts repeated for any of the participants. On average, each text was read 29 times.

These subjective ratings of text comprehension and reading speed are the foundation for the readability criteria used in this study. These ratings are different than traditional measures of comprehension such as multiple-choice questions and cloze test responses and were selected because they provide a more subjective benchmark for comprehension – that is, these ratings may measure comprehension more effectively when normed across 100s of raters and 100s of text. In addition, the approach allows for readers to judge not only text comprehension but also text reading speed, an element of readability that is not captured in multiple-choice and cloze test responses. Lastly, text comparisons such as these are simple and lend themselves to online collection techniques, such as crowdsourcing.

Calculating text difficulty

To calculate pairwise comparison scores for the two categories of participant text ratings (i.e., text comprehension and reading speed), we used a Bradley–Terry model (Bradley & Terry, 1952). Our modified version ranked documents by difficulty based on each text's probability to be more difficult than other texts, using either text comprehension or reading

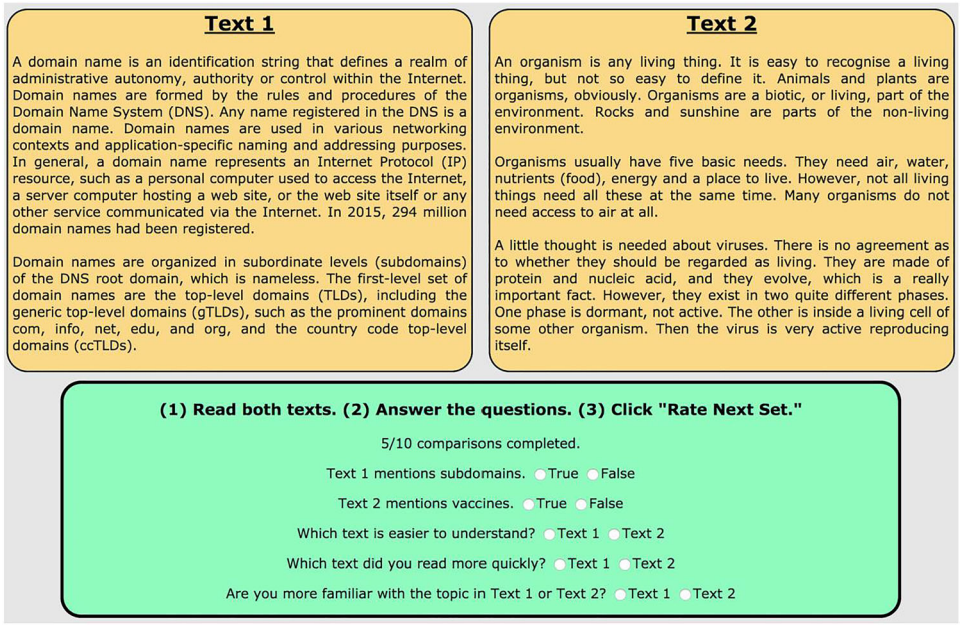

Figure 1. Example text rating screen seen by online participants. [Colour figure can be viewed at wileyonlinelibrary.com]

speed rankings. The model creates a maximum likelihood estimate that iteratively converges towards a unique maximum that defines the ranking of the texts (i.e., the most difficult ones have the highest probability). The probability values are then used as criteria (both text comprehension and reading speed) for developing readability formulas.

Linguistic feature selection

We used NLP tools to calculate linguistic features for each of the reading texts. The selected linguistic features overlap with text complexity measures associated with theories of reading and included text cohesion, lexical sophistication, sentiment analysis and syntactic complexity features. Lexical sophistication and semantic features were calculated using the Tool for the Automatic Analysis of Lexical Sophistication (Kyle & Crossley, 2015), ReaderBench (RB; Dascalu et al., 2014) and the Sentiment Analysis and Cognition Engine (SEANCE; Crossley et al., 2017). Syntactic complexity was calculated using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016), while text cohesion was calculated using the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley et al., 2016). In addition, we used RB to calculate traditional readability scores for each text (i.e., Flesch–Kincaid Grade Level, Flesch Reading Ease and the New Dale–Chall Readability Index). A more detailed account of these tools and the linguistic features they reported are provided in Dascalu (2014), Crossley et al. (2016), Crossley et al. (2017), Kyle (2016) and Kyle and Crossley (2015).

Tool for the Automatic Analysis of Lexical Sophistication. Tool for the Automatic Analysis of Lexical Sophistication (Kyle & Crossley, 2015) calculates over 200 indices related to the

lexical sophistication of words in a text. Measure of lexical sophistication correlates with faster processing and more accurate word decoding. Tool for the Automatic Analysis of Lexical Sophistication measures simple lexical information (i.e., number of word and n-gram types), word frequency (i.e., how many times a lexical item occurs in a larger, reference corpus), vocabulary range (i.e., how many documents in a reference corpus that a lexical item appears), psycholinguistic word information (e.g., familiarity, concreteness and meaningfulness), academic language (i.e., words and n-grams that occur frequently in academic writing and speech), lexical decision and naming times and accuracies, strength of association features, word neighbour information, contextual distinctiveness and semantic lexical relations such as polysemy (i.e., the number of senses a word has) and hypernymy (i.e., how specific a word is).

Frequency and range indices are calculated from available corpora such as the British Frequency and range indices are calculated fromNational Corpus (BNC Consortium; 2007), Brown corpus (Kučera & Francis, 1967) and the Corpus of Contemporary American English (Davies, 2009). Beyond frequency, bigram and trigram indices also measure proportion scores (i.e., the proportion of common n-grams found in a reference corpus). Psycholinguistic word information indices come from the Medical Research Council psycholinguistic database (Coltheart, 1981) or other freely available databases (e.g., Brysbaert, Warriner, & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). Contextual distinctiveness measures, which measure the diversity of contexts in which a word is encountered, are also taken from freely available databases (Adelman, Brown, & Quesada, 2006; Brysbaert & New, 2009; McDonald & Shillcock, 2001) as are word recognition and naming norms (Balota et al., 2007).

Sentiment Analysis and Cognition Engine. Sentiment Analysis and Cognition Engine (Crossley et al., 2017) is a sentiment analysis tool that relies on a number of pre-existing sentiment, social positioning and cognition dictionaries. SEANCE can provide semantic information about the words in a text, which may help readers link common themes and ideas, and SEANCE contains a number of predeveloped word vectors developed to measure sentiment, cognition and social order. These vectors are taken from freely available source databases. For many of these vectors, SEANCE also provides a negation feature (i.e., a contextual valence shifter) that ignores positive terms that are negated (e.g., not happy).

Tool for the Automatic Analysis of Syntactic Sophistication and Complexity. Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (Kyle, 2015; Kyle & Crossley, in press) measures both coarse and fined grained clausal and phrasal complexity, both of which are related to syntactic complexity. Greater syntactic complexity can make it more difficult to organise main ideas within a text and assign thematic roles, leading to difficulties in parsing sentences effectively (Graesser et al., 1996; Mesmer et al., 2012). At the coarse level, TAASSC reports on the 14 indices measured by Lu's Syntactic Complexity Analyser (Lu, 2010, 2011). At the more granular level, TAASSC reports on 31 indices of clausal complexity and 132 indices of phrasal complexity. The Syntactic Complexity Analyser indices calculate syntactic complexity based on t-unit analyses (Ortega, 2003; Wolfe-Quintero et al., 1998). The granular clausal indices calculate syntactic complexity based on the average incidence of particular structures per clause and dependents per clause. The granular phrasal indices calculate seven noun phrase types and 10 phrasal dependent types.

Tool for the Automatic Analysis of Cohesion. Tool for the Automatic Analysis of Cohesion (Crossley et al., 2016) reports on over 150 indices related to text cohesion. Sensitivity to cohesion structures in a text can help readers process and understand paragraphs and larger discourse segments (Gernsbacher, 1990; Mesmer et al., 2012). For many cohesion indices, TAACO integrates a part of speech tagger found in the Stanford Parser (Manning et al., 2014) and synonym sets reported by the WordNet lexical database (Miller, 1995). TAACO provides lexical overlap at the sentence (local cohesion) and paragraph (global cohesion) level for function words, content words and part of speech tags as well for synonym overlap. TAACO also calculates type token ratio (TTR) indices (for content words, function words and n-grams). Moreover, TAACO computes a variety of connective indices such as temporal connectives (e.g., *before* and *after*) and order connectives (e.g., *first* and *next*).

ReaderBench. In addition to reporting traditional readability formulas, RB (Dascalu, 2014) also reports on a number of surface level features and lexical features used in follow-up analyses and not reported in the other NLP tools used within this study. At the surface level, RB calculates features including average word and sentence length and average sentences per paragraph. Lexically, RB reports entropy indices for words distributions.

Classic readability formulas

All classic readability formulas were reported by RB and followed the formulas reported by Kincaid et al. (1975) and Flesch (1948) for reporting Flesch–Kincaid Grade Level and Flesch Reading Ease, respectively. For the New Dale–Chall Readability Formula, RB follows the computational implementation of the formula as reported by Crossley et al. (2017).

Statistical analysis

We first developed regression models to predict text comprehension and reading speed ratings. We initially removed any texts that were identified as outliers (above three standard deviations away from the mean) based on the pairwise comparison scores. This left us with 591 texts for the comprehension model and 595 texts for the text reading speed model. Prior to the regression analyses, we randomly split the data sets into training (~67% of the data) and test (~33% of the data) sets. We also removed any linguistic variables reported by the NLP tools that violated a normal distribution to better assure that residuals were normally distributed. In most cases, discarded variables represented linguistic features with low coverage that occurred extremely seldom in the data (and therefore were not candidates for transformation). Pearson correlations were then conducted on the remaining variables to determine whether the variables were meaningfully correlated with judgments of text complexity (i.e., text comprehension and reading speed). Any variables that did not reach an absolute correlation value of $r \geq .100$ with the text complexity judgments, which represents the threshold for a small effect (Cohen, 1988), were removed from further consideration. The remaining variables were checked for multicollinearity to ensure that the final model consisted only of indices that reported unique variance and that multicollinear indices would not reduce the likelihood that a potential variable would predict additional variance in the regression models. For each pair of variables with absolute correlation values of $r \geq .699$, only the variable with the highest correlation with text

complexity judgments was retained. In order to control for type 1 errors, we removed all variables that reported a $p < .001$.[3] The remaining variables were entered into a stepwise regression analysis using the data from the training set. The model derived from this analysis was afterwards extended to the remaining cases found in the test set in order to assess the generalisability of the model on held-out data.

Results

Crowdsourced algorithm of reading comprehension (CAREC)

We conducted correlations between the selected linguistic indices and the text comprehension ratings generated by the Bradley–Terry model for how easy a text was to understand. After controlling for normal distribution, effect sizes, multicollinearity and multiple comparisons, the text comprehension judgment analysis included 63 linguistic indices.

We conducted a stepwise regression analysis using the selected linguistic indices as the independent variables to analyse which linguistic features predicted the text comprehension ratings in the training set. This yielded a significant model, $F(13, 381) = 14.148$, $p < .001$, $r = .571$, $R^2 = .326$. Thirteen variables were significant predictors of the text comprehension ratings including variables related to lexical sophistication (word age of acquisition, frequency, imageability and character entropy), n-gram features (bigram range, trigram proportion scores and trigram TTR), cohesion (lexical overlap and the paragraph and sentence level) and sentiment (positive adjectives). Results from the regression are reported in Table 2. We used the *B* weights and the constant from the training set multiple regression analysis to estimate how the model would function on an independent data set (i.e., the test set). The model for the test set reported $r = .537$, $R^2 = .288$, demonstrating that the combination of the 13 variables accounted for 29% of the variance in the pairwise judgments found in the test set.

Crowdsourced algorithm of reading speed (CARES)

We conducted correlations between the selected linguistic indices and the text reading speed ratings generated by the Bradley–Terry model for how quickly a text was read. After controlling for normal distribution, effect sizes, multicollinearity and multiple comparisons, the reading speed judgment analysis included 51 linguistic indices.

We conducted a stepwise regression analysis using the selected linguistic indices as the independent variables to analyse which linguistic features predicted the reading speed ratings in the training set. This yielded a significant model, $F(9, 400) = 25.452$, $p < .001$, $r = .611$, $R^2 = .373$. Nine variables were significant predictors of the text comprehension ratings including variables related to text structure (number of content lemmas, number of function words and standard deviation for number of paragraphs), lexical sophistication (word naming responses times, word concreteness, semantic distinctiveness and characters per word) and syntactic complexity (complex nominals per t-unit and number of dependents per direct object nominal). Results from the regression are reported in Table 3. We used the *B* weights and the constant from the training set multiple regression to estimate how the model would function on an independent data set (i.e., the test set). The model for the test set reported $r = .649$, $R^2 = .421$, demonstrating that the combination of the nine variables accounted for 42% of the variance in the pairwise judgments found in the test set.

Table 2. Summary of multiple regression model for comprehension ratings (CAREC).

Entry	Predictors included	*r*	R^2	*β*	*SE*	*B*	*t*
1	Age of acquisition content words (Kuperman)	.391	.153	0.022	0.011	0.112	2.032*
2	Bigram range (COCA academic)	.427	.183	0.746	0.251	0.152	2.976**
3	Trigram proportion score (BNC written)	.461	.213	−0.742	0.184	−0.209	−4.029***
4	Trigram type token ratio	.482	.232	−0.699	0.210	−0.159	−3.334***
5	Lexical overlap (paragraph)	.495	.245	−0.111	0.048	−0.100	−2.311*
6	Characters per word (SD)	.504	.254	0.047	0.021	0.101	2.207*
7	Temporal connectives	.514	.264	−2.067	0.807	−0.110	−2.560*
8	Imageability content words (MRC)	.522	.273	−0.001	0.001	−0.175	−3.671***
9	Positive adjectives	.532	.283	−0.08	0.035	−0.104	−2.293*
10	Noun overlap (sentence)	.542	.293	0.035	0.012	0.140	2.944**
11	Lemma type count (content words)	.553	.306	0.002	0.001	0.157	3.285***
12	Character entropy	.563	.317	−0.395	0.135	−0.136	−2.916**
13	Word frequency (Brown)	.571	.326	0.000054	0.001	0.125	2.147*

Note. Constant = 1.811. CAREC, crowdsourced algorithm of reading comprehension; COCA, Corpus of Contemporary American English; MRC, Medical Research Council.
*$p < .050$.
**$p < .010$.
***$p < .001$.

Table 3. Summary of multiple regression model for reading speed ratings (CARES).

Entry	Predictors included	*r*	R^2	*β*	*SE*	*B*	*t*
1	Number of content words (lemmas)	.396	.157	0.004	0.001	0.388	7.646**
2	Word naming response time	.525	.276	0.003	0.001	0.150	2.855**
3	Complex nominals per t-unit	.551	.304	0.011	0.006	0.086	1.791*
4	Sentences per paragraph (SD)	.564	.318	−0.015	0.005	−0.123	−2.836**
5	Concreteness all words (MRC)	.576	.332	−0.001	0.001	−0.146	−3.155**
6	Semantic distinctiveness all words	.587	.345	−0.461	0.134	−0.172	−3.429***
7	Characters per word (SD)	.595	.354	0.062	0.021	0.127	2.934**
8	Number of function words	.604	.365	0.002	0.001	0.146	2.835**
9	Dependents per direct object nominal	.611	.373	0.023	0.01	0.091	2.221*

Note. Constant = −.862. CARES, crowdsourced algorithm of reading speed; MRC, Medical Research Council.
*$p < .050$.
**$p < .010$.
***$p < .001$.

Comparisons with traditional readability formulas

We used the crowdsourced models for comprehension (crowdsourced algorithm of reading comprehension [CAREC]) and reading speed (crowdsourced algorithm of reading speed [CARES]) to develop readability scores for each of the 600 texts in our Wikipedia corpus. For the same 600 texts, we calculated readability scores for Flesch Reading Ease, Flesch–Kincaid Grade Level and the New Dale–Chall Readability Index. We then calculated associations between the readability formulas using Pearson correlations. The correlations indicated strong effect sizes between the judgments of text comprehension and reading speed with CAREC and CARES, respectively. Medium and weak effect sizes were reported between traditional readability models and CAREC and CARES (see Table 4 for the correlation matrix). The New Dale–Chall reading formula reported the strongest correlations, while weaker correlations were reported for Flesch Reading Ease scores and Flesch–Kincaid Grade Level.

We used Fisher r-to-z transformation to examine differences in the correlation for the text comprehension and reading speed ratings for the crowdsourced models and the correlations reported by the traditional readability formulas. The r-to-z transformations demonstrated that CAREC and CARES predicted a significantly greater amount of variance than all the traditional readability formulas for the pairwise scores for text comprehension and reading speed. No differences were noted between the variance explained by the traditional readability formulas (see Tables 5 and 6 for details).

Discussion

The goal of this study was to explore how two models of readability (CAREC and CARES), based on linguistic features taken from state-of-the-art NLP tools, predicted text

Table 4. Correlations between crowdsourced ratings and each readability model/formula.

Variable	Crowdsourced models	Flesch Reading Ease	Flesch–Kincaid Grade Level	New Dale–Chall
Comprehension scores	0.560	−0.309	0.316	0.399
Reading speed scores	0.623	−0.296	0.309	0.355

Table 5. Fisher r-to-z transformation comparisons between readability formula: Text comprehension ratings.

Formulas	2	3	4
1. CAREC	5.370***	5.240***	3.610**
2. Flesch–Kincaid Reading Ease	–	0.130	1.770
3. Flesch Grade Level	–	–	1.630
4. New Dale–Chall Readability Index	–	–	–

CAREC, crowdsourced algorithm of reading comprehension.
*$p < .05$.
**$p < .010$.
***$p < .010$.

Table 6. Fisher r-to-z transformation comparisons between readability formula: Reading speed ratings.

Formulas	2	3	4
1. CARES	7.280***	7.040***	6.150***
2. Flesch Reading Ease	–	0.250	1.130
3. Flesch–Kincaid Grade Level	–	–	0.890
4. New Dale–Chall Readability Index	–	–	–

CARES, crowdsourced algorithm of reading speed.
*$p < .05$.
**$p < .010$.
***$p < .010$.

comprehension and reading speed measures collected using crowdsourcing techniques. We found that both CAREC and CARES outperformed classic readability formulas, thus providing support for the use of advanced NLP tools in the development of new formulas. In doing so, we also further report on the inherent weaknesses of classic readability models. Lastly, the linguistic features that inform CAREC and CARES indicate how text complexity indices can provide support for theoretical models of text readability.

CAREC explained 33% of the variance in crowdsourced comprehension ratings for the training set and 29% of the variance in the test set. The model included 13 variables related to lexical sophistication, n-gram features, text cohesion and sentiment. In terms of lexical sophistication, texts appear to be more difficult to comprehend when they have more content words, words learned later (i.e., higher age of acquisition), have more characters per word, are less imageable and have less character entropy (i.e., less diversity of characters). N-gram features indicated that texts with fewer common trigrams (as reported by proportion scores) and with less repetition of trigrams (as reported by TTR score) are more difficult to comprehend. In addition, texts that contain more general bigrams common in academic writing (as reported by range scores) are judged to be more difficult to comprehend. Text cohesion features suggested that passages with less lexical overlap at the paragraph level (i.e., less global cohesion) and fewer temporal connectives were more difficult to comprehend. Conversely, texts with greater local cohesion as evidenced by noun overlap between sentences were more difficult to process. In terms of affect/semantic, texts that contained fewer positive adjectives were more difficult to process. In total, these features present an overview of a text that is more difficult to comprehend as one that contains words that are more difficult to decode and provides less opportunity to develop meaning construction through the use of word overlap at the paragraph level and temporal connectives. These findings provide evidence that CAREC attests to many discourse processes that are important in readability theory (i.e., decoding and meaning construction, Just & Carpenter, 1980; Snow, 2002). Additionally, CAREC indicates that more positive texts are easier to read as are texts that have fewer content words.

Crowdsourced algorithm of reading speed explained 38% of the variance in crowdsourced reading speed ratings in the training set and 42% of the variance in the test set. The model supports theories of readability in a number of ways. First, texts that were more difficult to process (i.e., required more effort to read) contained more words, words that were sophisticated (i.e., words that take longer to name, are less concrete, less semantically diverse and contain more characters per word). Second, more difficult texts

contained sentences that were more syntactically complex (i.e., sentences that contained more complex t-units and more dependents per direct object). In addition, texts that required more effort to read had less variation in paragraph size and contained more function words. In total, these findings provide support for theories of readability because elements of discourse, specifically decoding and syntactic complexity (Just & Carpenter, 1980; Snow, 2002), are important predictors of reading speed. This finding differs slightly from CAREC that indicated that comprehension relies on decoding words and then integrating those words within a coherent model of meaning with the help of text cohesion features. While reading speed is predicted by variables related to decoding and syntactic parsing, it is not predicted by cohesion variables. This makes sense because cohesion is needed to develop knowledge representation important for comprehension, but it may not be related to reading speed. In addition, differences in function and content word counts were reported between CAREC and CARES. More content lemmas were predictive of greater comprehension difficulty, whereas more function words were predictive of greater reading speed. This finding is likely the result of function words assisting in the development of meaning construction by structuring relations between words during comprehension processes. However, during reading processes, a greater number of function words in a text may cause readers to spend more time processing a text as a result of a more elaborated discourse structure.

Beyond developing new models of text comprehension and reading speed, we also compared CAREC and CARES predication rates with the prediction rates of classic readability formulas. In each case, our crowdsourced models significantly outperformed the classic readability formulas in terms of amount of explained variance. Importantly, no classic readability formula performed significantly better than another. Thus, the crowdsourced models developed in this study are not only more theoretically aligned with text readability but they also outperform classic readability formulas in predicting crowdsourced ratings of text comprehension and reading speed. In addition, CAREC and CARES were developed on a large, modern corpus of everyday texts, and the text difficulty judgments were collected from a large and diverse population of adults, which likely affords greater generalisations towards larger populations of texts and readers.

An important aspect of this study is the use of crowdsourced judgments of comprehension and reading speed. While such judgments have been used in the past (De Clercq et al., 2014, Crossley et al., 2017), they represent an important new direction in the collection of reading criteria. Crowdsourced judgments allow for quick and efficient data collection at reasonable costs, allowing for much larger data sets to be collected. This is important because most classic readability formulas were based on very small sample sizes (generally under 100 texts). However, it is an open question as to whether subjective judgments of comprehension are better measures of comprehension than cloze tests or multiple-choice questions. There is reason to think judgments might be a better proxy of comprehension because they directly tap into readers experiences with a text, but it is unknown whether these judgments reflect deeper text comprehension. Pairwise comparisons can also capture judgments of reading speed, which cannot be measured using cloze tests or multiple-choice items. Again, it is an open question as to how accurately these reflect effortful processing on the part of readers. One problem with the current interface is that it was impossible to assess how long readers spent on each text because the texts needed to be displayed next to each other for the comparisons. Future studies involving eye-tracking or masked texts that are unmasked at mouse rollover may be able to address this concern.

Conclusion

This study developed two new readability formulas (CAREC and CARES) using NLP methods, to predict crowdsourced reading judgments of reading speed and text comprehension. These formulas, based on linguistic features that better represent theoretical and behavioural accounts of the reading process, significantly outperformed classic readability formulas. However, we do not have strong evidence about how well CAREC and CARES will generalise to other populations. The population tested here consisted mostly of adult readers with at least a high enough literacy to engage in on-line crowdsourcing tasks. Future research should test how well CAREC and CARES work for children and adolescents or how well the methods used here could be used to develop age specific reading models. Lastly, work is underway to make CAREC and CARES freely available in a user-friendly, downloadable tool so that researchers, teachers and administrators can receive automatic readability scores on texts and help lessen the reliance on readability formulas. Once available, the tool will be hosted on the website linguisticsanalysistools.org. Until then, we provide a tutorial in Appendix A on how to calculate the formulas using available NLP tools.

Notes

1. A t-unit is a dominant clause and its dependent clauses (Hunt, 1965).
2. The Dale–Chall formula is an exception in that it includes a list of frequent words.
3. After checking for normal distributions, muticollinearity and effect size, the number of available features dropped to ~50, which led to a Bonferroni α correction of .001.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determined word-naming and lexical decision times. *Psychological Science,* 17, 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x 16984300, 9

Allington, R.L. (2005). NCLB, reading first, and whither the future. *Reading Today*, 23(2), 18.

Bailin, A. & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3), 285–301. https://doi.org/10.1016/S0271-5309(01)00005-2.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B. et al. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459. https://doi.org/10.3758/BF03193014.

Begeny, J.C. & Greene, D.J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools*, 51(2), 198–215.

Benjamin, R.G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1), 63–88.

BNC Consortium. (2007). The British National Corpus, version 3. BNC consortium. Retrieved from www.natcorp.ox.ac.uk

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

Britton, B.K. & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83, 329–345.

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. https://doi.org/10.3758/BRM.41.4.977.

Brysbaert, M., Warriner, A.B. & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5.

Buhrmester, M., Kwang, T. & Gosling, S.D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3–5. https://doi.org/10.1177/1745691610393980.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd edn). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins-Thompson, K. & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the Association for Information Science and Technology*, 56(13), 1448–1462.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4), 497–505. https://doi.org/10.1080/14640748108400805.

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

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., 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., Kyle, K. & McNamara, D.S. (2017). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, 49(3), 803–821.

Crossley, S.A., Skalicky, S., Dascalu, M., McNamara, D.S. & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264.

Cunningham, J.W., Spadorcia, S.A., Erickson, K.A., Koppenhaver, D.A., Sturm, J.M., & Yoder, D.E. (2005). Investigating the instructional supportiveness of leveled texts. *Reading Research Quarterly*, 40(4), 410–427. https://doi.org/10.1598/RRQ.40.4.2

Dale, E. & Chall, J.S. (1949). The concept of readability. *Elementary English*, 26(1), 19–26.

Dascalu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating: Studies in computational intelligence*, Vol 534. Cham, Switzerland: Springer.

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

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159–190. https://doi.org/10.1075/ijcl.14.2.02dav.

Davison, A. & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2), 187–209.

De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M. & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering*, 20(3), 293–325.

DuBay, W.H. (2004). *The principles of readability.* Costa Mesa, CA: Impact Information.

Feng, L., Jansche, M., Huenerfauth, M. & Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, (pp. 276–284). USA: Association for computational linguistics.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

François, T. & Fairon, C. (2012, July). An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 466–477). USA: Association for Computational Linguistics.

François, T. & Miltsakaki, E. (2012, June). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, (pp. 49–57). USA: Association for Computational Linguistics.

Fry, E.B. (1989). Reading formulas: Maligned but valid. *Journal of Reading*, 32, 292–297.

Geiser, S. & Studley, W.R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, 8(1), 1–26.

Gernsbacher, M.A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Givón, T. (1995). *Functionalism and grammar.* Philadelphia: John Benjamins.

Goodman, J., Cryder, C. & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3), 213–224. https://doi.org/10.1002/bdm.1753.

Graesser, A.C., McNamara, D.S. & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223–234.

Graesser, A.C., Swamer, S.S., Baggett, W.B. & Sell, M.A. (1996). New models of deep comprehension. In B.K. Britton & A.C. Graesser (Eds.), *Models of understanding text*, (pp. 1–32). Mahwah, NJ: Erlbaum.

Grimes, J.E. (1975). *The thread of discourse*. The Hague, Netherlands: Mouton.

Heilman, M., Collins-Thompson, K., Callan, J. & Eskenazi, M. (2006). Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *9th International conference on spoken language processing*. Pittsburgh, PA: ISCA.

Hiebert, E.H. (2011). Beyond single readability measures: Using multiple sources of information in establishing text complexity. *Journal of Education*, 191(2), 33–42.

Howes, D.H. & Solomon, R.L. (1951). Visual duration thresholds as a function of word probability. *Journal of Experimental Psychology*, 41(6), 401–410.

Hunt, K. W. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3.

Juel, C. & Solso, R.L. (1981). The role of orthographic redundancy, versatility and spelling-sound correspondences in word identif ication. (1981). In M.L. Kamil (Ed.), *Directions in reading: Research and instruction*, (pp. 74–82). Rochester, NY: National Reading Conference.

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. et al. (2010, August). Learning to predict readability using diverse linguistic features. In *In Proceedings of the 23rd International Conference on Computational Linguistics*, (pp. 546–554). USA: Association for Computational Linguistics.

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. (No. RBR-8-75)*. Naval Technical Training Command, Millington, TN: Research Branch.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163–182. https://doi.org/10.1037/0033-295X.95.2.163.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, 49, 294–303. https://doi.org/10.1037/0003-066X.49.4.294.

Kintsch, W., Welsch, D., Schmalhofer, F. & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, 29, 133–159. https://doi.org/10.1016/0749-596X(90)90069-C.

Klare, G.R. (1984). Readability. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (Vol. 1, pp. 681–744). New York: Longman.

Kučera, H. & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44, 978–990. https://doi.org/10.3758/s13428-012-0210-4.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. (Doctoral dissertation), Georgia State University. Retrieved from http://scholarworks.gsu.edu/alesl_diss/35

Kyle, K. & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49, 757–786. https://doi.org/10.1002/tesq.194.

Kyle, K. & Crossley, S.A. (in press). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Language Journal*, 102(2), 333–349.

Litman, L., Robinson, J. & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioural sciences. *Behavior Research Methods.*, 49(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15, 474–496. https://doi.org/10.1075/ijcl.15.4.02lu.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45, 36–62. https://doi.org/10.5054/tq.2011.240859.

Magliano, J.P., Millis, K., Ozuru, Y. & McNamara, D.S. (2007). A multidimensional framework to evaluate reading assessment tools. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies*, (pp. 107–136). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. (2014). The Stanford CoreNLP natural language processing toolkit In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

McDonald, S.A. & Shillcock, R.C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44, 295–323.

McNamara, D.S. & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22, 247–288. https://doi.org/10.1080/01638539609544975.

Mesmer, H.A. (2005). Decodable text and the first grade reader. *Reading & Writing Quarterly*, 21(1), 61–86. https://doi.org/10.1080/10573560590523667.

Mesmer, H.A., Cunningham, J.W. & Elfrieda, H.H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, 47(3), 235–258.

Mesmer, H.A., Cunningham, J.W., & Hiebert, E.H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, 47(3), 235–258.

Miller, G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41.

Miller, J.R. & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 335–354. https://doi.org/10.1037/0278-7393.6.4.335.

National Assessment of Educational Progress. (2011). The Nation's Report Card: Writing 2011.

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

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24, 492–518. https://doi.org/10.1093/applin/24.4.492.

Pitler, E. & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 186–195). USA: Association for Computational Linguistics.

Powell, P.R. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication*, 60, 664–682.

Richards, J.C., Platt, J. & Platt, H. (1992). *Longman dictionary of language teaching and applied linguistics*. London: Longman.

Richardson, J.T.E. (1975). The effect of word imageability in acquired dyslexia. *Neuropsychologia*, 13(3), 281–288.

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 523–530).

Snow, C. (Ed.) (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. Santa Monica, CA: Rand.

Stanovich, K.E. (1985). Explaining the variance in reading ability in terms of psychological processes: What have we learned? *Annals of Dyslexia*, 35(1), 67.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (2003). *The nation's report card: Reading highlights 2003 (NCES 2004–452)*. Washington, DC: National Center for Education Statistics.

Wikipedia. (n.d.). About Wikipedia. Retreived from https://simple.wikipedia.org/wiki/Main_Page

Wolfe, M.B., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P.W., Kintsch, W. et al. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, 25(2–3), 309–336.

Wolfe-Quintero, K., Inagaki, S. & Kim, H.-Y. (1998). In University of Hawai'i, Second Language Teaching and Curriculum Center (Ed.), *Second language development in writing: Measures of fluency, accuracy, and complexity.* Honolulu, HI.

Zamanian, M. & Heydari, P. (2012). Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1), 43.

Appendix A

This appendix contains instructions for calculating the Crowdsourced Algorithm of REading Comprehension (CAREC) and the Crowdsourced Algorithm of REading Speed (CARES). The models rely on five freely available NLP tools that are listed below (along with websites for downloads). We are currently working on a single NLP tool that will calculate both CAREC and CARES.

- Tool for the Automatic Analysis of Cohesion (TAACO version 1.5.2; linguisticanalysistools.org)
- Tool for the Automatic Analysis of Lexical Sophistication (TAALES version 2.2; linguisticanalysistools.org)
- Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC version 1.1; linguisticanalysistools.org)
- Sentiment Analysis and Cognition Engine (SEANCE version 1.2.0; linguisticanalysistools.org)
- ReaderBench (http://www.readerbench.com)

Currently, to derive both CAREC and CARES, texts need to be in .txt format and then need to be run through each of the tools listed above. Each tool will produce a .csv spreadsheet that can be opened using software such as Microsoft Excel. Each text will be a row and each column in that row will be a calculation for a linguistic feature found in the text.

The tables below (Tables 1 and 2) show which indices need to be selected from each tool along with their labels in the tools' output along with their co-efficient values in the regression model. The co-efficients are needed to calculate a texts' predicted comprehension score or reading speed.

To calculate the predicted comprehension of text, the reported scores for each of the indices needs to be derived from the tools above for that text. A relatively simple algorithm is needed to get the final score. The algorithm involves adding the constant for each model (CAREC = 1.811, CARES = -0.862) to each reported linguistic feature score from the text multiplied by the co-efficient for that score in the model. Partial examples for CAREC and CARES are provided below. Full examples are given in Tables 1 and 2.

CAREC = 1.811 + (Kuperman_AoA_CW value * 0.022) + (COCA_Academic_Bigram_Range Value * 0.746) + (BNC_Written_Trigram_Proportion value * -0.742) + remaining features and co-efficients.

CARES = -0.862 + (nlemma_content_words value * 0.004) + (WN_Mean_RT value * 0.003) + (CN_T * 0.011) + remaining features and co-efficients.

Dr Scott Crossley is a professor of Applied Linguistics at Georgia State University. Professor Crossley's primary research focus is on natural language processing and the application of computational tools and machine learning algorithms in language learning, educational success, writing and text comprehensibility.

Stephen Skalicky is a Lecturer of Applied Linguistics in the School of Linguistics and Applied Language Studies at Victoria University of Wellington. Stephen's primary research focus is on quantitative investigation of investigate links among language, cognition, and creativity.

Mihai Dascalu holds a double PhD in Computer Science and Educational Sciences. He has extensive experience in national and international research projects and is a Professor at the Politehnica University of Bucharest where he teaches courses on Object Oriented Programming, Semantic Web Applications, and Data Mining and Data Warehousing.

Received 3 August 2018; revised version received 4 July 2019.

Address for correspondence: Scott A. Crossley, Department of Applied Linguistics/ESL, 34 Peachtree St. Suite 1200, One Park Tower Building, Georgia State University, Atlanta, GA 30303, USA. E-mail: *scrossley@gsu.edu*