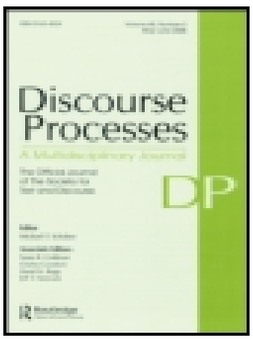

Discourse Processes

ISSN: 0163-853X (Print) 1532-6950 (Online) Journal homepage: http://www.tandfonline.com/loi/hdsp20

Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas

Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara & Kristopher Kyle

To cite this article: Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara & Kristopher Kyle (2017): Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas, Discourse Processes, DOI: 10.1080/0163853X.2017.1296264

To link to this article: http://dx.doi.org/10.1080/0163853X.2017.1296264

Published online: 20 Mar 2017.

DISCOURSE PROCESSES
http://dx.doi.org/10.1080/0163853X.2017.1296264

Routledge
Taylor & Francis Group

Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas

Scott A. Crossley[a,b], Stephen Skalicky[a,b], Mihai Dascalu[a,b], Danielle S. McNamara[c], and Kristopher Kyle[d]

[a]Department of Applied Linguistics/ESLGeorgia State University; [b]Department of Computer Sciences Politehnica University of Bucharest, Bucharest, Romania; [c]Department of Psychology Learning Sciences Institute Arizona State University; [d]Department of Second Language Studies University of Hawaii, Manoa

ABSTRACT

Research has identified a number of linguistic features that influence the reading comprehension of young readers; yet, less is known about whether and how these findings extend to adult readers. This study examines text comprehension, processing, and familiarity judgment provided by adult readers using a number of different approaches (i.e., natural language processing, crowd-sourced ratings, and machine learning). The primary focus is on the identification of the linguistic features that predict adult text readability judgments, and how these features perform when compared to traditional text readability formulas such as the Flesch-Kincaid grade level formula. The results indicate the traditional readability formulas are less predictive than models of text comprehension, processing, and familiarity derived from advanced natural language processing tools.

Introduction

Literacy is an important component not only of educational success in primary, secondary, and postsecondary contexts, but also success in business and in life (Geiser & Studley, 2002; Powell 2009). However, becoming literate is a long, complex, and difficult undertaking that requires the coordination of a number of cognitive and knowledge skills (National Assessment of Educational Progress, 2011). As text becomes more common in mediums such as e-mails, text messages, and social media posts, the need for strong literacy skills will continue to increase (Barton & Lee, 2013; National Assessment Governing Board, 2011; National Writing Project, 2010). The majority of literacy data reports on school-aged readers, with grim indications that students struggle to develop strong reading skills (Carnegie Council on Advancing Adolescent Literacy, 2010; Grigg, Daane, Jin, & Campbell, 2003). For instance, 25% or more of students in the 8th and 12th grades in the United States perform below a basic level of reading comprehension (U.S. Department of Education, 2011). These students have an increased risk of being referred to special education classes, missing grade level advancements, and dropping out of school (Reynolds & Ou, 2004; Shonkoff & Phillips, 2000). Literacy problems continue after secondary schooling, with over 25% of entering college students enrolled in remedial classes (U.S. Department of Education, 2003). Reading demands increase in the workplace (Pennsylvania Department of Education, 2004), and many adults in the United States may not be ready for the challenge (Goodman, Finnegan, Mohadjer, Krenzke, & Hogan, 2013; Kutner, Greenburg, Jin, & Paulsen, 2006).

One way to support and scaffold literacy challenges is to match text difficulty to the abilities of the reader. Providing students with texts that are accessible and well matched to their reading abilities

CONTACT Scott A. Crossley scrossley@gsu.edu Department of Applied Linguistics/ESL, 25 Park Place, Suite 1500, Georgia State University, Atlanta, GA 30303, USA.

Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/hdsp

helps to ensure that they better understand the text and, across time, helps readers to improve their reading skills. Indeed, since the 1970s more than 200 readability formulas have been produced in the hopes of providing tools to measure text difficulty more accurately and efficaciously. Most of these readability formulas are based on factors that represent two broad features of text difficulty: lexical sophistication and syntactic complexity. These features are generally measured by word and sentence length, respectively. Many available readability formulas are not based on any particular theory of reading or reading comprehension and are instead based on empirical correlations. Therefore, their soundness is strictly predictive and they are often accused of having weak construct validity (i.e., they do not actually measure what they purport to measure; Davison & Kantor, 1982). Nonetheless, a number of classic validation studies have found the formulas' predictive validity to be consistently high, correlating with observed difficulty in the range of .7 to .8 (Chall, 1958; Chall & Dale, 1995; Fry, 1989). Such high correlations may be attributable to the use of cloze testing to develop reading criteria (Bormuth, 1966) and/or to the testing of readability formulas on simplified texts (i.e., texts that have been modified to ease comprehension on the part of the reader, especially for beginning level readers). The use of cloze testing is problematic because it is strongly associated with word and sentence length (Crossley, Dufty, McCarthy, & McNamara, 2007), the very features used by many readability formulas to predict difficulty. In a similar fashion, developing readability formulas for simplified text is problematic because simplified texts are generally modified to contain more frequent (familiar) words and shorter sentences (Chall & Dale, 1995; Crossley, Allen, & McNamara, 2011, 2012).

Nevertheless, the draw toward readability formulas' simple, mechanical assessments has led to their widespread use by teachers, administrators, testing agencies, print media, and the military to find reading material that is appropriate for different levels of readers (DuBay, 2004), including materials designed for a wide variety of readers and reading situations beyond those for which the formulas were created. For instance, readability formulas are used to assess authentic texts (as compared with simplified or modified texts), nonacademic texts, and texts designed for non-student populations. The widespread use of traditional formulas in spite of their restricted validity has inclined many researchers within the field of discourse processing to regard them with reservation (Bruce & Rubin, 1988; Bruce, Rubin, & Starr, 1981; Davison & Kantor, 1982; Rubin, 1985; Smith, 2012). Additionally, the rise of cognitive models of reading has underscored not only the limitations of the traditional formulas but also the need for measures that account for discourse-specific factors, such as text-based and situation level processing (Kintsch, Welsch, Schmalhofer, & Zimny, 1990; McNamara, Kintsch, Butler Songer, & Kintsch, 1996), as well as individual reader variables, such as background knowledge and reading skill. A more inclusive assessment of text comprehensibility must go deeper than surface readability features to explain how a learner interacts with a text (Kintsch, 1994; McNamara et al., 1996; Miller & Kintsch, 1980).

Achieving a more inclusive assessment of text comprehension requires collecting comprehension performance on multiple levels (e.g., reading time, text-based questions, situation model questions) across multiple genres and multiple populations, as well as individual difference measures such as reading skill and prior knowledge. This would require resources beyond those available to most (or any) researchers and is of little interest to those who profit from the commercialization of traditional readability formulas. Hence, alternative approaches must be examined to further our understanding of text difficulty and the multiple factors that influence it.

In this study we introduce such an alternative approach, one that uses crowd sourcing techniques to collect human judgments of text comprehension, processing, and familiarity and develops reading criteria for these judgments using statistical pairwise comparisons between readers' judgments. We then use sophisticated natural language processing (NLP) tools to assess an extensive range of linguistic features in the texts that go beyond lexical and syntactic complexity. We use these linguistic features to predict the pairwise ratings for each text, in the process deriving models of text comprehension, processing, and familiarity. We then compare the predictive strengths of these models to traditional readability formulas.

Readability formulas

We discuss a number of common readability formulas below, including Flesch Reading Ease (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975), and Dale-Chall (Chall & Dale, 1995).

Flesch formulas. The Flesch Reading Ease formula (Flesch, 1948) is probably the earliest readability formula to have a widespread impact on text development and selection. The criterion for the formula was the average grade level of a child who could correctly answer three-fourths of test questions asked about a given passage in the McCall-Crabbs' Standard test lessons in reading (1926). The formula itself was based on sentence length and number of syllables per word. The reported correlation between the formula and the reading criterion was .71. A number of follow-up studies using parenting texts and texts intended for remedial adult readers reported correlations between .55 and 87 (Klare, 1952). In the 1970s the Flesch Reading Ease formula was recalculated using the same text features on 18 passages taken from Navy training manuals (Kincaid et al., 1975). No correlations with reading criteria were reported for the new formula called the Flesch-Kincaid Grade Level formula. However, it produced higher intercorrelations compared with two other formulas: the Automated Readability Index (Senter & Smith, 1967) and the Gunning Fog Index Readability (FOG) formula (Gunning, 1952).

Dale-Chall formulas. The original Dale-Chall formula was created in 1948 to address a number of issues involving affix counts and personal references included in an earlier version of the Flesch formula (Flesch, 1943). The Dale-Chall formula also differed in that it used a large list of frequent words to assess vocabulary difficulty (n = 3,000). The original Dale-Chall formula was based on sentence length and the number of words in the text that were not found in the list of common words. The correlation between this formula and the McCall-Crabbs' Standard test lessons in reading criterion was .70. Chall improved upon the formula (Chall & Dale, 1995) by including an updated list of frequent words. Chall and Dale tested the new formula on the 32 passages reported by Bormuth (1971) and reported a correlation of .92 for both the new and old Dale-Chall formulas.

Criticisms of readability formulas. A number of criticisms have been leveled at traditional readability formulas. First and foremost, the formulas are criticized for not having strong construct validity. This is because the formulas are generally not based on theories of reading or comprehension but rather rely on statistical correlations to develop predictive power. For instance, most theories of text comprehension suggest that text readability is based on a number of component features that operate at different levels of linguistic processing, including lexical, syntactic, semantic, and discourse features (Just & Carpenter, 1980; Koda, 2005). However, readability formulas are based on only two levels of linguistic features (i.e., lexical and syntactic), and these features are, at best, proxies of the features recognized as important during linguistic processing. These formulas do give a rough estimate of difficulty, but because many are based simply on the number words in a sentence and the number of letters per word, the formulas will judge the readability of nonsense texts to the same degree as authentic texts (Davison & Kantor, 1982). Those formulas that do use word frequency lists often use small lists that may or may not be updated and may or may not include technical terms, alternative forms of words (e.g., lemmas), or proper nouns (Collins-Thompson & Callan, 2004; Schwarm & Ostendorf, 2005).

In addition, traditional readability formulas do not take into consideration relationships between elements in the text (i.e., text cohesion), which help readers build new knowledge and are important indicators of text readability (Britton & Gülgöz, 1991; Kintsch, 1988; McNamara et al., 1996; McNamara & Kintsch, 1996), although this notion is complicated by background knowledge, with high-knowledge readers benefiting from texts that are lower in cohesion (McNamara, 2001; McNamara et al., 1996; O'Reilly & McNamara, 2007). Traditional readability formulas also do not consider style, vocabulary, and grammar, all of which may play an important role in readability (Bailin & Grafstein, 2001). More

importantly, text representations in the minds of a reader are not based on linguistic features alone but also include world knowledge, an important predictor of text readability (McNamara & Kintsch, 1996). Moreover, a very basic notion of what text readability encompasses is not widely defined. Most text readability formulas are based on text comprehension scores, but these scores are usually based on multiple choice questions and cloze tests, which may not serve as valid measures of deep levels of text comprehension (Magliano, Millis, Ozuru, & McNamara, 2007). In addition, some readability formulas are validated based on their ability to predict text level and not aspects of reading comprehension. Finally, readability likely encompasses some notion of text processing (i.e., how fast a text can be read, not just comprehended), but this notion is under researched in terms of readability formulas.

Although traditional readability formulas are still in in widespread use today, advances in NLP have created the potential to modernize readability formulas using more innovative and conceptually valid linguistic techniques (Benjamin, 2012). For instance, contemporary NLP tools report on a number of text features that have strong overlap with theoretical accounts of the reading process (Crossley, Greenfield, & McNamara, 2008). For instance, NLP tools can measure a variety of lexical features related to text decoding that go well beyond the length of words. These features include frequency counts; lexical properties related to word concreteness, meaningfulness, familiarity, and imageability; and word response latencies, all of which are theoretically important components in understanding word processing (Balota et al., 2007; Brysbaert, Lagrou & Stevens, in press; Brysbaert & New, 2009; Kuperman, Drieghe, Keuleers, & Brysbaert, 2013). In addition, modern NLP tools can measure text cohesion, which is a critical element in helping readers build text knowledge (McNamara et al., 1996). Such cohesion is measured at both the local (sentence) and global (paragraph) level by assessing lexical and semantic overlap between various text segments and the use of connectives and causal particles and verbs throughout a text.

A number of studies have started to harness the power of these NLP tools. For instance, Schwarm and Ostendorf (2005) developed a formula that included traditional measures along with measures of syntactic complexity (e.g., number of noun and verb phrases, parse height, and embedded clauses). They used this formula to successfully predict text reading level (second through fifth grades). Heilman, Collins-Thompson, Callan, and Eskenazi (2006) used the frequency of common grammatical constructions as a measure of grammatical difficulty to predict the grade level of texts and found that the inclusion of grammatical features lowered the error rate in level classification. Crossley et al. used the NLP tool, Coh-Metrix, to develop readability formulas for L1 (2007) and L2 (2008) readers that included measures of syntactic complexity, word frequency, and text cohesion and found that the formulas performed as well as traditional readability formulas with L1 readers (i.e., it reached a ceiling effect on a small corpus of texts judged for readability) and outperformed traditional readability formulas with L2 readers. In a similar fashion, Pitler and Nenkova (2008) combined lexical, syntactic, and discourse features to predict judgments of readability. They used a small, nonacademic corpus (The Wall Street Journal corpus) and found that linguistic features related to syntax, semantics, and discourse were strong predictors of readability, whereas traditional readability formulas were not.

These more recent studies provide indications of how corpus linguistics, NLPtools, and machine learning can be used to develop newer readability formulas that have greater overlap with theories of reading and more strongly tap into linguistic features in the text. However, these studies are limited in their generalizability because they depended on small corpora (Crossley et al., 2007; Pitler & Nenkova, 2008), the use of texts that were likely developed based on readability formulas (Crossley et al., 2007), or predicted grade level but not text readability (Heilman et al., 2006; Schwarm & Ostendorf, 2005).

Current study

The current study used crowdsourcing techniques to collect human judgments of text comprehension, processing, and familiarity. From these judgments, we used a Bradley-Terry model, which calculated pairwise comparisons among the ratings to estimate the difficulty of each text in comparison to the other texts. Using linguistic features taken from NLP tools, we examined the potential to develop models for judgments of text comprehension, processing, and familiarity. We

compared the accuracy of these models to classic readability formulas (e.g., Dale-Chall and Flesch Reading Ease). Unlike previous research, we used crowdsourcing techniques to gather human judgments of text comprehension, processing, and familiarity. We also used a number of state-of-the-art NLP tools to derive our linguistic features. Our research questions were as follows:

(1) What are the relationships between judgments of text comprehension, processing, and familiarity?
(2) Are linguistic features predictive of judgments of text comprehension, processing, and familiarity?
(3) Are the derived models of text difficulty more predictive than models derived from classic text readability formulas?

Methods

Corpus

The texts used for this study were nonacademic news articles selected from the Guardian Weekly, a British-based publication with a wide international readership. In total, 150 news articles were selected. One hundred of these articles were from a corpus of simplified texts used in previous studies (Allen, 2009; Crossley et al., 2011, 2012). These 100 texts were taken from an English teaching website (www.onestopenglish.com), which provides simplified news texts and accompanying learning activities for second language learners of English. The articles were simplified by a small, independent team of authors into three levels of simplification: advanced, intermediate, and beginning. For this study, 50 beginning and 50 intermediate articles were randomly selected. In addition, 50 original news articles were selected. Text selection was controlled for topic across levels so that a similar number of texts on different topics (such as business, culture, environment, media, politics, science, and world news) were found at each level. We selected beginning, intermediate, and original texts to ensure variance in the difficulty of the texts used. We also controlled for text length by truncating texts so they were approximately 150 words. Texts were truncated at the end of the articles using natural paragraph breaks.

Participants

Participants were recruited using the crowdsourcing service Mechanical Turk available through Amazon.com. Amazon Mechanical Turk (AMT) is a well-established crowdsourcing service in which workers anonymously complete short online tasks in return for small fees. We used crowdsourcing techniques to collect text difficulty judgments because the data could be collected efficiently and with less expense. The AMT population is also more diverse and just as reliable as traditional university undergraduate research participant pools. At the same time, the AMT worker population possesses homogenous qualities that can provide greater generalizability of results (Buhrmester, Kwang, & Gosling, 2011; Goodman, Cryder, & Cheema, 2013), which has not the case with previous studies of readability. We used a third-party website (TurkPrime; Litman, Robinson, & Abberbock, 2016) to recruit participants because it helps block duplicate workers and manages the development and deployment of website experiments.

English speaking AMT workers from the United States were offered $1.50 to participate. Those who agreed to participate were briefed on the nature and purpose of the experiments and then provided informed consent. In total, 307 participants were recruited and completed the study. Eight participants were removed for either not answering at least 75% of comprehension questions correctly (see below) or reading the texts too quickly (<3 standard deviations of mean). In total, 299 participants provided 3011[1] pairwise comparisons from which we derived text comprehension, text processing, and text familiarity ratings for each text.

[1]Because of errors in the website, two participants completed a combined total of 17 extra ratings and three participants completed fewer than 10 ratings, resulting in a total that does not equal 10 ratings per participant.

Table 1. Participant survey information.

Hours of TV		Books per Year		Reading Confidence		Reading Enjoyment	
Response	Percent	Response	Percent	Response	Percent	Response	Percent
None	5.686	None	5.351	Very unconfident	1.672	Not at all	1.003
<1 hour	17.391	1 book	8.361	Unconfident	0.669	Not very much	2.676
1 hour	19.398	2 books	7.692	Somewhat unconfident	1.003	Somewhat	22.742
2 hours	26.421	3 books	13.043	Somewhat confident	11.706	Quite a bit	30.100
3 hours	18.395	4 books	9.030	Confident	35.786	Very much	43.478
4 hours	6.020	5 books	7.692	Very confident	49.164		
5 hours	3.010	5 or more	48.829				
6 hours	2.341						
7 hours	0.669						
≥8 hours	0.669						

Study design

We used custom JavaScript and HTML code to design a website from which to collect readability judgments. The website first collected survey information from participants about their first language, how many books they read a year, how many hours of television they watched each day, how much they enjoyed reading, and how confident they were in their reading ability. Seven participants were not first language speakers of English. Moat participants self-reported that they watched around 1 to 3 hours of television a day, read about five or more books a year, and were confident readers that enjoyed reading (see Table 1 for survey results).

After the survey, participants were then asked to make 10 text comparisons. During these comparisons, participants saw two texts side by side. Participants were asked to read the texts and then answer one true or false comprehension question per text. The comprehension questions focused on the main ideas of the text and were used to confirm that participants had read each text and were on task. The questions were not used as reading criteria. Participants then had to complete three forced ratings in which they selected which text was easier to understand (i.e., text comprehension), which text they read more quickly (i.e., text processing), and with which text they were more familiar (i.e., knowledge of topic). See Figure 1 for an overview of the interface. The comparisons, aggregated over multiple participants and texts, allowed us to evaluate the

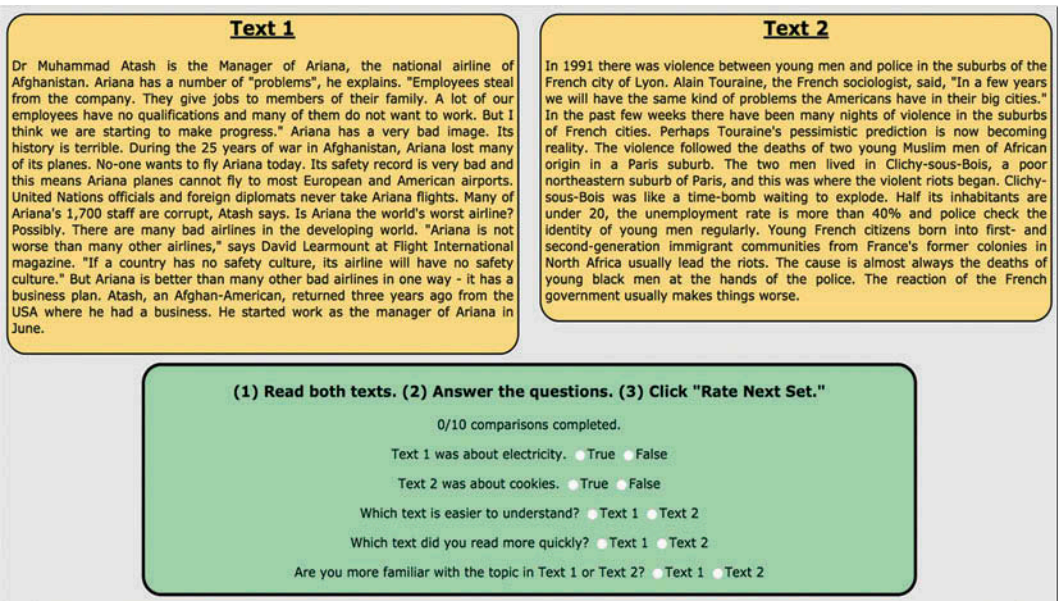

Figure 1. Study interface.

difficulty of texts in comparison with one another, forming the readability criteria used in this analysis. Texts were randomly paired and participants saw each text only once. Each text was read 40 times on average.

Calculating text difficulty

We used a Bradley-Terry model (Bradley & Terry, 1952) to conduct pairwise comparisons among the participant ratings for the comprehension, text processing, and familiarity ratings. Individual models were trained for each of the previous three ranking criteria. A Bradley-Terry model describes the probabilities of the possible outcomes when individuals are judged against one another in pairs (see Eq. 1). Over multiple comparisons, our modified Bradley-Terry model evaluates a text's difficulty as its likelihood to be more difficult than another text based on a criterion of interest. To create the most plausible ranking of documents, we introduced a randomly initialized, normalized vector γ, in which γi is a positive-valued parameter associated with document i, for each of the comparisons in terms of the difficulty on a certain criterion between texti against textj.

$$P\left(text_i \text{ more difficult than } text_j\right) = \frac{\gamma_i}{\gamma_i + \gamma_j} \tag{1}$$

Linguistic features of text

We used a number of NLP tools to derive linguistic features from the reading texts. These linguistic features measured lexical sophistication, text cohesion, and syntactic complexity, and were reported by four tools: The Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, 2016), the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015), the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC; Kyle, 2016), and ReaderBench (RB; Dascalu, Dessus, Trausan-Matu, Bianco, & Nardy, 2013a). In addition, we calculated traditional readability scores for the texts (i.e., Flesch-Kincaid Grade Level, Flesch Reading Ease, and the New Dale-Chall Readability Index). These readability scores and indices were used to predict the ratings for text comprehension, processing, and familiarity calculated by the Bradley-Terry model. We briefly discuss the indices and readability scores below. More detailed overviews of these tools are provided in Dascalu (2014), Crossley et al. (2016), Kyle (2015), and Kyle and Crossley (2015).

Tool for the Automatic Analysis of Cohesion. TAACO (Crossley et al., 2016) incorporates over 150 classic and recently developed indices related to text cohesion. For a number of indices, the tool incorporates a part of speech tagger from the Natural Language Tool Kit (Bird, Klein, & Loper, 2009) and synonym sets from the WordNet lexical database (Miller, 1995). TAACO provides linguistic counts for both sentence and paragraph markers of cohesion and incorporates WordNet synonym sets. Specifically, TAACO calculates type token ratio indices (for all words, content words, function words, and n-grams), sentence overlap indices that assess local cohesion for all words, content words, function words, part of speech tags (e.g., nouns, verbs, adjectives, and adverbs), and synonyms, paragraph overlap indices that assess global cohesion for all words, content words, function words, part of speech tags, and synonyms, and a variety of connective indices such as logical connectives (e.g., moreover, nevertheless), conjuncts (however, furthermore), the incidence of and, and order connectives (e.g., first, before, after).

Tool for the Automatic Analysis of Lexical Sophistication. TAALES (Kyle & Crossley, 2015) calculates approximately 200 indices related to basic lexical information (i.e., the number of words and n-grams, the number of word and n-gram types), lexical frequency (i.e., how many times a word occurs in a reference corpus), lexical range (i.e., how many documents in which a reference corpus an item occurs), psycholinguistic word information (e.g., concreteness, familiarity, meaningfulness),

academic language (i.e., items that occur more frequently in an academic corpus than in a general use corpus) for both single words and multiword units (e.g., n-grams such as bigrams and trigrams), strength of association, contextual distinctiveness, word neighbor information, lexical decision times, age of exposure, and semantic lexical relations such as hypernymy (i.e., word specificity) and polysemy (i.e., word ambiguity).

The frequency and range indices draw on the British National Corpus (2007), Thorndike-Lorge Corpus (Thorndike & Lorge, 1944), Brown corpus (Kučera & Francis, 1967), Brown verbal frequencies (Brown, 1984), which were compiled based on the London-Lund corpus of English Conversation (Svartvik & Quirk, 1980), the SUBTLEXus corpus of subtitles (Brysbaert & New, 2009; Davies, 2009), and the Corpus of Contemporary American English (COCA; Davies, 2009). Bigram and trigram indices include frequency and proportion scores (i.e., the proportion of common n-grams found in a reference corpus). Psycholinguistic word information indices draw on the Medical Research Council psycholinguistic database (Coltheart, 1981), which includes word scores for familiarity, concreteness, imageability, and meaningfulness (how many associations a word has). Also included in TAALES are psycholinguistic word information indices that incorporate recently collected concreteness norms for single words and two-word units (Brysbaert, Warriner, & Kuperman, 2014) and age of acquisition norms for single words (i.e., at what age a word is estimated to be learned; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). Academic language frequencies are based on the Academic Word List (Coxhead, 2000) and the Academic Formulas List (Simpson-Vlach & Ellis, 2010). TAALES also calculates five association measures for each bigram and trigram found in COCA: Mutual Information, Mutual Information Squared, t-score, ΔP, and collexeme score. Mutual Information, Mutual Information Squared,t-score, and collexeme score are bidirectional measures of association between constituent words in an n-gram, whiereas ΔP scores are unidirectional. TAALES also calculates a number of indices related to contextual distinctiveness approach that measure the diversity of contexts in which a word is encountered (Adelman, Brown, & Quesada, 2006; Brysbaert & New, 2009; McDonald & Shillcock, 2001). TAALES also incorporates measures from The English Lexicon Project, a large publicly available psycholinguistic dataset (Balota et al., 2007). The English Lexicon Project indices include word recognition norms (e.g., response latencies and accuracies for word recognition and lexical naming tasks) and word neighborhood information (e.g., the number of words that are in a words orthographic neighborhood). Finally, TAALES incorporates age of exposure metrics, which reflect the age/grade level of a learner and are based on computational models that estimate a word's complexity based on co-occurrence data (Dascalu, McNamara, Crossley, & Trausan-Matu, 2016).

Tool for the Automatic Analysis of Syntactic Sophistication and Complexity. TAASSC (Kyle, 2015) measures large and fined grained clausal and phrasal indices of syntactic complexity and usage-based frequency/contingency indices of syntactic sophistication. TAASSC includes 14 indices measured by Lu's (2010, 2011) Syntactic Complexity Analyzer, 31 fine-grained indices or clausal complexity, 132 fine-grained indices of phrasal complexity, and 190 usage-based indices of syntactic sophistication. The Syntactic Complexity Analyzer measures are classic measures of syntax based on t-unit analyses (Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). The fine-grained clausal indices calculate the average number of particular structures per clause and dependents per clause. The fine-grained phrasal indices measure seven noun phrase types and ten phrasal dependent types. The syntactic sophistication indices are grounded in usage-based theories of language acquisition (Ellis, 2002; Goldberg, 1995; Langacker, 1987) and measure the frequency, type token ratio, attested items, and association strengths for verb-argument constructions in a text.

ReaderBench. RB (Dascalu, 2014) is a multilingual framework that integrates a multitude of textual complexity indices ranging from classic readability formulas, surface indices, syntactic features, as well as semantics and discourse features (i.e., text cohesion and connectivity). RB integrates multiple semantic models including Latent Semantic Analysis (Landauer, & Dumais, 1997), Latent Dirichlet

Allocation (Blei, & Lafferty, 2009) and semantic distances in WordNet, namely Wu-Palmer, Leakock-Chodorow, and path length (Budanitsky & Hirst, 2006). At the surface level, RB reports on predictive linguistic proxies such as average word and sentence length, average unique/content words per text, and average commas per sentence. RB also calculates entropy indices at the word level (e.g., the expected occurrence of characters within a word). At the syntax level, RB quantifies textual complexity in terms of syntax structure using different part of speech tags including nouns, verbs, prepositions, adjectives and adverbs.

In addition, RB calculates the number of semantic dependencies or depth of a parsed sentence, which can indicate a more complex structure of discourse. Semantically, RB uses local and global evaluations of cohesion within a Cohesion Network Analysis graph to compute the average value of the semantic similarities of all links within a text (Dascalu et al., 2013a; Dascalu, Trausan-Matu, & Dessus, 2013b) at intra- and interparagraph levels. In addition, RB calculates discourse centered indices based on the interanimation of voices. These voices are operationalized as semantic chains of interconnected concepts whose distribution and overlap define the structure of discourse. RB uses Pointwise Mutual Information to best captures the synergic effect between voices in terms of their co-occurrences (Dascalu et al., 2013b). RB also uses an entropy measure with voices to measure the diversity of used concepts.

RB also computes additional features related to discourse connectedness based on lexical chains. These include the average and maximum span of lexical chains (the distance in words between the first and the last words pertaining to the same chain), average number of lexical chains per paragraphs, and percentage of words that are included in lexical chains and that are not isolated within the discourse. At the semantic level RB calculates the incidence of general nouns and named entities (e.g., people's names, venues, organizations), which have an important role in text comprehension because established entities form the basic components of concepts. Finally, RB reports on a number of word complexity features including syllable count, distance in characters between the inflected form, lemma and word stem counts (where adding multiple prefixes or suffixes increases the difficulty of using a certain word), distinguishability reflected in the inverse document frequency from a text corpus (in this case, the Touchstone Applied Science Associates corpus), and the distance in a hypernym tree (i.e., how specific a word is).

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

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

Flesch Reading Ease. We calculated the Flesch Reading Ease based on the formula reported by Flesch (1948). Like Flesch-Kincaid Grade Level, this formula is based on the number of words per sentence (sentence length) and the number of syllables per word (word length). This formula is Flesch Reading Ease = – (1.015 × number of words/number of sentences) – (84.600 × number of syllables/number of words) + 206.835.

New Dale-Chall readability formula. The Dale-Chall readability formula (Chall & Dale, 1995) was designed to be hand-coded on exact samples of 100 words. In addition, the formula has a number of different rules for lemmatizing the words in its familiar word list. For instance, common contractions such as "it's" and "haven't" were included in the familiar word list, but users were instructed to also count familiar words that were in possessive, plural, past tense, present participle, third-person singular, comparative, and superlative forms. No mention was made of other inflectional forms (i.e., past participle) and users were instructed to not count derivational morphemes (e.g., "-tion", "-ation", "-ment", "-ly", "-y") as familiar. Other rules were given for compounds and hyphenated words, numerals, abbreviations, and names and places. Our computational automatization of the formula included an extended familiar word list that

included all possessives, plurals, past tense, present participle, third-person singular, comparative, and superlative forms for the original words on the list (n = 8,775) and followed the rules for other word forms. We did not limit our count to the first 100 words because the formula was automated. The formula is New Dale-Chall readability formula = 0.1579 × (difficult words/words × 100) + 0.0496 × (number of words/number of sentences).

Statistical analysis

We first examined correlations among the judgments for text comprehension, processing, and familiarity along with the traditional readability formulas. We then developed regression models to predict text comprehension, processing, and familiarity ratings. Before this analysis we removed any variables that violated a normal distribution to better assure that residuals were distributed normally. In most cases, discarded variables represented linguistic features that occurred extremely rarely in the data (and therefore were not candidates for transformation). Pearson correlations were then conducted on the remaining variables to determine whether they were meaningfully correlated with judgments of text complexity (i.e., the comprehension, processing, and familiarity ratings). 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. To control for Type 1 errors, we removed all variables that reported a $p > .001$. The remaining variables were checked for multicollinearity to ensure that the final model consisted only of unique indices and that multicollinear indices did not exaggerate the results of the multiple regression analysis (Tabachnick, Fidell, & Osterlind, 2012). For each pair of variables with absolute correlation values of $r \geq .899$, only the variable with the highest correlation with text complexity judgments was retained.

The remaining variables were entered into a stepwise regression analysis followed by a 10-fold cross-validation multiple regression using the selected variables. Ten-fold cross-validation is a method designed to avoid overfitting a statistical or machine-learning model (Witten, Frank, & Hall, 2011). In a 10-fold cross-validation multiple regression, the dataset is randomly divided into 10 sections (called "folds"). A stepwise multiple regression is conducted using 9 of the 10 folds to train a statistical model, which is then tested on the remaining fold. This procedure is repeated nine more times until all of the folds have served as the test set and each of the 10 models is averaged.

Results

Correlations among text judgments and traditional readability formulas

Pearson correlations indicated medium to strong effect sizes among the judgments of text comprehension, processing, and familiarity (see Table 2 for the correlation matrix). Specifically, judgments of text comprehension were strongly correlated with judgments of text processing and familiarity. Judgments of text processing and familiarity showed a medium effect size with each other. In terms of traditional readability formulas, the Dale-Chall formula moderately correlated with judgments of text processing and comprehension, with a weak correlation with text familiarity judgments. The Flesch Reading Ease and Flesch-Kincaid

Table 2. Correlations among text judgments and traditional readability formulas.

Measure	2	3	4	5	6
1. Text processing judgments	.681[a]	.488[a]	−.328[a]	.385[a]	.426[a]
2. Text comprehension judgments	—	.679[a]	−.238[a]	.285[a]	.329[a]
3. Text familiarity judgments	—	—	−0.14	0.133	.213[a]
4. Flesch Reading Ease	—	—	—	−.974[a]	−.592[a]
5. Flesch-Kincaid grade level	—	—	—	—	.657[a]
6. New Dale-Chall	—	—	—	—	—

[a] $p < .01$

formulas moderately correlated with judgments of text processing, with very weak to nonsignificant correlations with judgments of text familiarity.

Judgments of text comprehension

We conducted correlations between the selected linguistic indices and the text comprehension ratings based on how easy a text was to understand generated by the Bradley-Terry model. After controlling for normal distribution, effect sizes, multicollinearity, and multiple comparisons, the text comprehension judgement analysis included 26 linguistic indices. Correlations are presented in Table 3.

To analyze which linguistic features predicted the text comprehension ratings, we conducted a stepwise regression analysis using the selected linguistic indices as the independent variables. This yielded a significant model, $F(4, 145) = 16.334$, $p < .001$, $R = .557$, $R2 = .331$. Four variables were significant predictors of the text comprehension ratings: Age of acquisition (Kuperman), Lexical decision accuracy scores, Content word overlap (adjacent paragraphs), and Average number of verbs. Only Lexical decision accuracy rating was a negative predictor (Table 4). The four variables used in a 10-fold cross-validation yielded R = .431, R2 = .186, indicating that the four variables together explained 19% of the variance in the text comprehension ratings.

Comparisons between models. We used Fisher r-to-z transformation to assess the significance of the differences between the correlation reported for the regression model and the traditional readability formulas (i.e., the New Dale-Chall Readability Formula, Flesch Reading Ease, and the Flesch-Kincaid Grade Level correlations reported in Table 2) for the comprehension ratings. The z transformations demonstrated that the regression model predicted a greater amount of variance than all of the traditional readability formulas. No differences were noted between the variance explained by the traditional readability formulas (see Table 5 for details).

Table 3. Correlations between selected indices and text comprehension ratings.

Index	Tool	*r*	*p*
Age of acquisition (Kuperman)	TAALES	0.376	<.001
Lexical decision mean RT scores	TAALES	0.370	<.001
Lexical decision accuracy scores	TAALES	−0.361	<.001
Average number of nouns	TAASSC	0.336	<.001
Range scores content words (COCA)	TAALES	−0.332	<.001
Average sentence voice mutual information	RB	0.331	<.001
Familiarity (Medical Research Council, content words)	TAALES	−0.329	<.001
Lexical naming RT scores	TAALES	0.326	<.001
Range scores (SUBTLEXus)	TAALES	−0.315	<.001
Average number of verbs	TAASSC	0.298	<.001
Average paragraph entropy of voices	RB	0.296	<.001
Frequency content words (COCA)	TAALES	−0.293	<.001
Average parsing tree size	RB	0.292	<.001
Average paragraph voice mutual information	RB	0.292	<.001
Age of exposure score	TAALES	0.287	<.001
Average sentence-paragraph cohesion (Inverse path length from WordNet)	RB	0.286	<.001
Verb phrases per T-unit	TAASSC	0.285	<.001
Lemma overlap (adjacent paragraphs)	TAACO	0.283	<.001
Bi-gram proportion (COCA fiction)	TAALES	−0.282	<.001
Number of entities per sentence	RB	0.281	<.001
Lemma overlap (adjacent sentences)	TAACO	0.281	<.001
Content word overlap (adjacent sentences)	TAACO	0.280	<.001
Order connectives	TAACO	0.274	<.001
Mean length of t-units	TAASSC	0.274	<.001
Average number of preposition	TAASSC	0.267	<.001
Word meaningfulness (all words)	TAALES	−0.267	<.001
Content word overlap (adjacent paragraphs)	TAACO	0.263	<.001
Noun lemma overlap (adjacent paragraphs)	TAACO	0.258	<.001

RB = ReaderBench.

Table 4. Summary of multiple regression model for comprehension ratings.

Entry	Predictors Included	*r*	R^2	R^2 change	β	*SE*	*B*
1	Age of acquisition (Kuperman)	0.376	0.141	0.141	0.002	0.001	0.184
2	Lexical decision accuracy scores	0.457	0.209	0.068	–0.414	0.097	–0.312
3	Content word overlap (adjacent paragraphs)	0.526	0.277	0.068	0.043	0.011	0.263
4	Average number of verbs	0.557	0.311	0.034	0.001	0.001	0.200

Constant = 0.379

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

Formulas	1	2	3
1. Current regression	2.46[a]	2.88[a]	3.31[b]
2. New Dale-Chall Readability Index	—	0.42	0.85
3. Flesch grade level	—	—	0.43
4. Flesch-Kincaid reading ease	—	—	—

[a] $p < .05$
[b] $p < .01$

Judgments of text processing

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 judgement analysis included 27 linguistic indices. Correlations are presented in Table 6.

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(3, 146) = 42.459, p < .001, R = .683, R2 = .466. Three variables were significant predictors of the text comprehension ratings: Trigram type count, Age of Acquisition (Kuperman), and Average number entities per sentence. All were positive predictors (Table 7). The three variables used in a 10-fold cross-validation yielded R = .661, R2 = .437, indicating that the three variables together explained 44% of the variance in the text processing ratings.

Comparisons between models. We used Fisher r-to-z transformation to assess differences in the amount of variance predicted for the text processing ratings between the regression model and the traditional readability formulas. The z transformations demonstrated that the regression model predicted a greater amount of variance than all the traditional readability formulas. No differences were noted in between the variance explained by the traditional readability formulas (see Table 8 for details).

Judgments of text familiarity

We conducted correlations between the selected linguistic indices and the text familiarity ratings generated by the Bradley-Terry model. After controlling for normal distribution, effect sizes, multicollinearity, and multiple comparisons, the text familiarity judgement analysis included six linguistic indices. Correlations are presented in Table 9.

To analyze which linguistic features predicted the text familiarity ratings, we conducted a stepwise regression analysis using the selected linguistic indices as the independent variables. This yielded a significant model, F(4, 145) = 12.519, p < .001, R = .507, R2 = .257. Four variables were significant predictors of the text familiarity ratings: Incidence of order words, Word frequency (COCA magazine), Paragraph overlap content words, and Lexical decision response times. All were positive predictors (Table 10). The fours variables used in a 10-fold cross-validation yielded R = .409, R2 = .167, indicating that the three variables together explained 17% of the variance in the text familiarity ratings.

Table 6. Correlations between selected indices and text processing ratings.

Index	Tool	*r*	*p*
Number of content words (lemma)	TAACO	0.479	<.001
Number of trigram types	TAACO	0.449	<.001
Average number of nouns per sentence	RB	0.417	<.001
Average sentence voice mutual information	RB	0.405	<.001
Age of acquisition (Kuperman)	TAALES	0.404	<.001
Word frequency (COCA fiction content words)	TAALES	–0.398	<.001
Bigram proportion (COCA fiction)	TAALES	–0.397	<.001
Average paragraph entropy of voices (content words)	RB	0.373	<.001
Average word length	TAALES	0.372	<.001
Range (SUBTLEXus)	TAALES	–0.355	<.001
Percentage of words that are included in lexical chains	RB	–0.350	<.001
Lexical decision response time	TAALES	0.348	<.001
Word name response times	TAALES	0.344	<.001
Number of words in orthographic neighborhood per word	TAALES	–0.341	<.001
Average number of entities per Rangesentence	RB	0.338	<.001
Average paragraph voice mutual information	RB	0.338	<.001
Range (COCA magazine)	TAALES	–0.331	<.001
Number of complex nominals per t-unit	TAASSC	0.329	<.001
Average number of verbs per sentence	TAASSC	0.299	<.001
Average number of prepositions per sentence	TAASSC	0.298	<.001
Verb phrase per t-unit	TAASSC	–0.297	<.001
Word familiarity	TAALES	–0.288	<.001
Age of exposure for words	TAALES	0.286	<.001
Bigram frequency (COCA fiction)	TAALES	0.283	<.001
Average sentence-paragraph cohesion	RB	0.279	<.001
Number of subjects per clause	TAASSC	–0.276	<.001
Average nominal dependents	TAASSC	0.271	<.001
Verb hypernymy	TAALES	0.263	<.001

Table 7. Summary of multiple regression model for pairwise comparisons (text processing).

Entry	Predictors Included	*r*	R^2	R^2 change	β	*SE*	*B*
1	Trigram type count	0.449	0.202	0.202	0.001	0.001	0.520
2	Age of acquisition (Kuperman)	0.651	0.424	0.222	0.006	0.001	0.389
3	Average number of entities per sentence	0.683	0.466	0.042	0.001	0.001	0.222

Constant = –0.060

Table 8. Fisher r-to-z transformation comparisons between readability formula: Text processing ratings.

Formulas	2	3	4
1. New regression	3.26[a]	3.68[a]	4.24[b]
2. New Dale-Chall Readability Index	—	0.42	0.98
3. Flesch Grade Level	—	—	0.56
4. Flesch-Kincaid Reading Ease	—	—	—

[a] $p < .05$
[b] $p < .01$

Table 9. Correlations between selected indices and text familiarity ratings.

Index	Tool	*r*	*p*
Incidence of order connectives	TAACO	0.330	<.001
Paragraph overlap content words	TAACO	0.296	<.001
Frequency (COCA magazine)	TAALES	–0.284	<.001
Bigram proportion scores (COCA magazine)	TAALES	–0.281	<.001
Paragraph overlap nouns	TAACO	0.276	<.001
Lexical decision response time	TAALES	–0.263	<.001

Table 10. Summary of multiple regression model for pairwise comparisons (text familiarity).

Entry	Predictors Included	*r*	R^2	R^2 change	β	*SE*	*B*
1	Incidence of order connectives	.330	0.109	0.109	0.23	0.07	0.243
2	Frequency (COCA magazine)	.418	0.174	0.066	−0.007	0.003	−0.188
3	Paragraph overlap content words	.480	0.231	0.056	0.038	0.011	0.257
4	Lexical decision response time	.507	0.257	0.026	−0.209	0.093	−0.178

Constant = 0.216

Table 11. Fisher r-to-z transformation comparisons between readability formula: Text familiarity ratings.

Formulas	2	3	4
1. New regression	2.94[a]	5.94[b]	6.00[b]
2. New Dale-Chall Readability Index	—	0.71	0.65
3. Flesch Grade Level	—	—	0.6
4. Flesch-Kincaid Reading Ease			

[a] $p < .05$
[b] $p < .01$

Comparisons between models. We used Fisher r-to-z transformation to examine differences in the prediction strength for the familiarity ratings for the regression model and the correlations reported for the traditional readability formulas. The z transformations demonstrated the regression model predicted a greater amount of variance than all the traditional readability formulas. No differences were noted in between the variance explained by the traditional readability formulas (see Table 11 for details).

Discussion

A number of readability formulas have been published in the last 70 years. However, most of these formulas have not undergone rigorous testing on text samples that have not been modified to increase reading comprehension (i.e., text samples designed for children and young readers). In addition, most formulas were not designed to assess readability in nonstudent populations, and most readability formulas are not based on theories of reading and thus lack strong construct validity. Indeed, most readability formulas indirectly assess two elements of text difficulty (lexical and syntactic complexity) and ignore other features of texts (including discourse features) and the reader (i.e., reading ability and background knowledge). These limitations have not stopped readability formulas from being widely adopted for use by teachers, administrators, testing agencies, print media, and the military to find and/or develop appropriate reading material (DuBay, 2004). The purpose of this study was to examine how predictive traditional readability formulas were in a reading task that examined adult readers' judgments of text comprehension, processing, and familiarity in a corpus of nonacademic texts. The results indicate the traditional readability formulas, when assessed on adult judgments of text comprehension on a representative corpus, are weaker indicators of text comprehension than previously reported. Traditional readability formulas were also less predictive than models of text comprehension, processing, and familiarity derived from advanced NLP tools.

Perhaps the strongest finding from this study is that readability formulas performed poorly in predicting adult readers' judgments of text comprehension. Early studies reported that traditional readability formulas explained around 50% of text comprehension for children in the case of the Flesch Reading Ease (Flesch, 1948) and around 80% of the variance in the case of the Dale-Chall formulas (Chall & Dale, 1995). In the case of the adult reading criteria collected for this study, the Dale-Chall formula explained around 10% of the variance of the human judgments, whereas the Flesch Reading Ease explained around 6% of the variance. A number of individual indices related to lexical sophistication and text cohesion outperformed these traditional readability formulas, and a combination of these linguistic features statistically outperformed the formulas. These indices

and the regression models reported from these indices support theoretical and psycholinguistic models of reading and indicate that decoding and meaning construction are important components of text comprehension. Specifically, the model demonstrates that texts are judged to be more comprehensive if the words are less sophisticated (i.e., lower age of acquisition scores and higher lexical decision accuracy score), there are fewer verbs, and lower text cohesion. Generally, greater text cohesion is thought to lead to more readable texts, but for competent readers, which many of the participants in this study likely were according to the survey results, high text cohesion does not aid in comprehension (McNamara, 2001; McNamara et al., 1996; McNamara & Kintsch, 1996; O'Reilly & McNamara, 2007).

This study also examined how linguistic features in the text related to human judgments of how quickly a text is processed. Although text processing time is an important component of understanding reading, few, if any, readability formulas have been developed to predict text processing speed and few, if any, studies have examined how well traditional readability formulas predict judgments of processing time. Although the human judgments of comprehension and processing were strongly related in this study (reporting an r = .681), about 50% of the variance in each criteria is not explained by the other. The results from this study indicate the Dale-Chall Readability Formula is a strong predictor of text processing ratings, explaining about 18% of the variance in the pairwise comparison scores. However, it is a weaker predictor than a number of individual linguistic features (Table 5). Flesch Reading Ease and Flesch Grade Level performed worse, explaining about 10% and 15% of the variance, respectively. When linguistic features that are linked to theoretical and psycholinguistic accounts of reading are combined, they statistically outperform all traditional readability formulas in explaining human judgments of text processing time. Specifically, texts that are judged to be processed slower contain a greater number of unique trigrams, have words with greater age of acquisition scores and a greater number of entities per sentence (i.e., proper nouns per sentence).

This study also investigated links between human judgments of text familiarity and text comprehension and processing and also developed NLP models to predict familiarity judgments and then compared these models to traditional readability formulas. The correlations in Table 2 indicate that judgments of text familiarity are strong predictors of judgments of comprehension and processing. Only the Dale-Chall formula was predictive of familiarity ratings, but the effect size for this correlation was small. Unlike the previous analysis of comprehension and processing ratings, only a few linguistic variables showed significant correlations after controlling for Type 1 errors. Of these variables, four were significant predictors in a regression analysis. These variables indicated that texts that were judged to be more familiar included fewer order words, more frequent words that had lower lexical decision response times, and less text cohesion.

Comparisons among models provide information about how linguistic features interact with human judgments of text complexity. For instance, comprehension judgments were strongly predicted by indices related to decoding (i.e., age of acquisition and lexical decision rates), discourse features (content word overlap), and the number of verbs. Similarities are found between the comprehension model and the processing model in that the processing model was also strongly informed by indices related to decoding (i.e., tri-gram type count and word age of acquisition). However, judgments of processing are not predicted by discourse features or number of verbs. Instead, the number of named entities is predictive of processing judgments such that a greater number of proper nouns slows processing. Intuitively this makes sense, because proper nouns will not be automatically decoded unless they are highly frequent. Greater overlap between the comprehension and familiarity ratings was reported in that both included indices related to cohesion and decoding. Theoretically this makes sense because text comprehension is a function of background knowledge, which may be captured by judgments of text familiarity (Kintsch, 1994; McNamara et al., 1996). It is also important to note that syntactic complexity indices were not included as predictors in any of our regression models. In fact, no syntactic complexity measures showed at least a medium correlation (.30) with human judgments of text comprehension with the strongest correlation

reported for the average parse tree size (.29) followed by the number of verb phrases per t-unit and the mean length of t-units. In terms of judgments of text processing, stronger correlations were reported. For instance, the number of complex nominals was a medium predictor of text processing ratings (correlation of .33) as was the number of verb phrase per t-unit (correlation of –.30). But neither of these were included in the regression analysis. No syntactic complexity indices were significant predictors of text familiarity. These findings indicate that syntactic complexity may not be strongly related to human judgments of comprehension, which counters findings from previous studies. For instance, both Pitler and Nenkova (2008) and Schwarm and Ostendorf (2005) reported that parse height and embedded clauses were predictive of text comprehension. It may be the case that syntax is not as important in developing models for human judgments of comprehension as lexical sophistication and/or text cohesion, potentially indicating that syntax does not interact with meaning construction in the same manner that lexical items and text cohesion do. However, behavioral studies are needed to test these assumptions. It may also be the case that syntax is an important predictor of objective comprehension measures but not the subjective comprehension measures used in this study.

This study also introduces a new approach for readability formulas: reader judgments. Traditional criteria used in the development of reading criteria included objective measures such multiple choice questions and cloze tests, both of which may not be adequate measures of text comprehension and are difficult to collect. Multiple choice questions always allow the potential for the right answer to be guessed and the questions themselves may prime the reader to recall information. In addition, multiple choice questions generally assess surface level knowledge and not deeper knowledge of a text such as inference generation. Cloze tests are problematic as well because longer sentences will increase the potential that multiple words will be removed from a single sentence. In addition, more frequent words are more likely to be deleted from a text. Thus, cloze tests results may favor linguistic indices related to sentence length and frequency (Crossley et al., 2007), the very features found in traditional readability formulas. Human judgments of text difficult avoid many of these drawbacks and are easier to collect, especially when crowdsourcing techniques are used. They are not flawless, of course, because human judgments can be skewed, whereas objective measures of readability can be more reliable. However, they serve to avoid many potential problems found in previous readability criterion and provide an alternative to objective measures.

Conclusion

This study has demonstrated that traditional readability formulas perform poorly in modeling adult judgments of text complexity in non-academic texts. A number of linguistic indices with strong overlap to theoretical and behavioral approaches to readability were more strongly predictive of text complexity. These findings raise concerns about the use of traditional readability formulas for populations and text types on which they were not normed. In contrast, the findings raise the potential for developing better readability formulas based on more advanced NLP indices, crowdsourcing techniques, and pairwise comparison algorithms.

However, the approaches and techniques used in this study need to be tested on a greater number of texts in a variety of domains and for a greater variety of readers. In addition, larger corpora need to be collected to ensure the results reported here are reliable across texts. It is also important to compare the pairwise comparison method used here with more traditional metrics of text readability such as multiple choice and cloze tests and, potentially, with more robust methods such as text recall and open-ended questions. This would allow for direct comparisons between subjective and objective measures. Once this is complete, the formulas need to be made publicly available so that practitioners and researchers can access the formulas and use them for developing texts and calculating text difficulty. In all cases the results reported in this study support concerns about traditional readability formulas and introduce new methods and approaches that may address these concerns and help with the development of more reliable readability formula.

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.

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

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, *21*, 285–301.

Bakhtin, M.M. (1981). *The dialogic imagination: Four essays* (C. Emerson & M. Holquist, Trans.). Austin and London: The University of Texas Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459.

Barton, D., & Lee, C. (2013). *Language online: Investigating digital texts and practices*. New York, NY: Routledge.

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

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. Sebastopol, CA: O'Reilly Media, Inc.

Blei, D. M., & Lafferty, J. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71–93). London, UK: Chapman & Hall/CRC.

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly*, *1*, 79–132.

Bormuth, J. R. (1971). Development of standards of readability: Toward a rational criterion of passage performance. Final report. Retrieved from http://files.eric.ed.gov/fulltext/ED054233.pdf.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*, 13–47.

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.

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

Bruce, B., & Rubin, A. (1988). Readability formulas: Matching tool and task. In G. M. G. Davison (Ed.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 5–22). Hillsdale, NJ: Erlbaum.

Bruce, B. C., Rubin, A. D., & Starr, K. S. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, *1*, 50–52.

Brysbaert, M., Lagrou, E., & Stevens, M. (in press). Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. *Bilingualism: Language and Cognition*, 1–19.

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*, 977–990.

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

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*, 3–5.

Chall, J. S. (1958). *Readability: An appraisal of research and application*. Columbus, OH: Ohio State University.

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

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

Collins-Thompson, K., & Callan, J. (2004). Information retrieval for language tutoring: An overview of the REAP project. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 544–545). New York, NY: ACM.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*, 497–505.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (Eds.). (2009). *Introduction to Algorithms* (3rd ed.). Cambridge, MA: MIT Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*, 86–101.

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research*, *16*, 89–108.

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

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*, 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*, 1227–1237.
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., Trausan-Matu, S., Bianco, M., & Nardy, A. (2013a). *ReaderBench, an environment for analyzing text complexity and reading strategies*. Presented at the 16th Int. Conf. on Artificial Intelligence in Education (AIED 2013), July 9–13, 2013, Memphis, TN.
Dascalu, M., McNamara, D. S., Crossley, S. A., & Trausan-Matu, S. (2016). Age of exposure: A model of word learning. In *30th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence* (pp. 2928–2934). Phoenix, AZ: AAAI Press.
Dascalu, M., Trausan-Matu, S., & Dessus, P. (2013b). Voices' inter-animation detection with ReaderBench—Modelling and assessing polyphony in CSCL chats as voice synergy. In *2nd Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 2nd Int. Conf. on Systems and* Computer Science *(ICSCS)* (pp. 280–285). Villeneuve d'Ascq, France: IEEE.
Dascalu, M., Trausan-Matu, S., Dessus, P., & McNamara, D. S. (2015). Dialogism: A framework for CSCL and a signature of collaboration. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *11th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2015)* (pp. 86–93). Gothenburg, Sweden: ISLS.
Dascalu, M., Trausan-Matu, S., & Dessus, P. (2013). Cohesion-based analysis of CSCL conversations: Holistic and individual perspectives. In N. Rummel, M. Kapur, M. Nathan, & S. Puntambekar (Eds.), *10th Int. Conf. on Computer-Supported Collaborative Learning (CSCL 2013)* (Vol. 1, pp. 145–152). Madison, WI: ISLS.
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*, 159–190.
Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, *17*, 187–209.
DuBay, W. H. (2004). *The principles of readability*. Costa Mesa, CA: Impact Information.
Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143–188.
Flesch, R. (1943). *Marks of readable style: A study in adult education* (Vol. 897). New York, NY: Columbia University Press.
Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.
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*, 1–26.
Goldberg, A. E. (1995). *Constructions: A construction grammar account of argument structure*. Chicago, IL: University of Chicago Press.
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*, 213–224.
Goodman, M., Finnegan, R., Mohadjer, L., Krenzke, T., & Hogan, J. (2013). Literacy, numeracy, and problem solving in technology-rich environments among US adults: Results from the program for the international assessment of adult competencies 2012. First Look. NCES 2014-008. Washington, D.C.: U.S. Department of Education, *National Center for Education Statistics* (NCES 2014-008). Retrieved from http://nces.ed.gov/pubsearch.
Grigg, W. S., Daane, M. C., Jin, Y., & Campbell, J. R. (2003). The nation's report card: Reading, 2002. Washington, D. C.: U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences (NCES 2003–531). Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2003521. *National Center for Education Statistics*.
Gunning, R. (1952). *The technique of clear writing*. New York, NY: McGraw-Hill.
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.
Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354.
Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability Formulas: (Automated readability index, fog count and Flesch Reading Ease Formula) for Navy enlisted personnel*. (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.
Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*, 294–303.
Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, *29*, 133–159.
Klare, G. R. (1952). Measures of the readability of written communication: An evaluation. *Journal of Educational Psychology*, *43*, 385–399.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge, MA: Cambridge University Press.
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.
Kuperman, V., Drieghe, D., Keuleers, E., & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *Quarterly Journal of Experimental Psychology*, *66*, 563–580.
Kutner, M., Greenburg, E., Jin, Y., & Paulsen, C. (2006). The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy (NCES 2006–483). Washington, D.C.: Government Printing Office.
Kyle, K. (2015). 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.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: Thelatent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211–240.
Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford, CA: Stanford University Press.
Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 1–10. doi: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.
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.
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.
McCall W. A. & Crabbs L. M. (1926). *Test Lessons in Reading Book S (Practice Lessons for Grades 5, 6, or 7)*. Teachers-College, Columbia University, New York, NY.
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. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, *55*, 51–62.
McNamara, D. S., Kintsch, E., Butler Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1–43.
McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–288.
Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 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.
O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43*, 121–152.
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.
Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 186–195). Honolulu, HI: Association for Computational Linguistics.
Powell, P. R. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication*, *60*, 664–682.
Reynolds, A. J., & Ou, S.-R. (2004). Alterable predictors of child well-being in the Chicago longitudinal study. *Children and Youth Services Review*, *26*, 1–14.
Rubin, A. (1985). How useful are readability formulas? In J. Osborn, P.T. Wilson, & R.C. Anderson (Eds.), *Reading Education: Foundations for a Literate America*, (pp. 61–77). Lexington, MA: Lexington Books.
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). Ann Arbor, MI: ACL.
Senter, R. J., & Smith, E. A. (1967). Automated readability index. Retrieved from www.dtic.mil/cgi-bin/GetTRDoc?AD=AD0667273.

Shonkoff, J. P., & Phillips, D. A. (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington, DC: National Academies Press.
Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*, 487–512.
Smith, F. (2012). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. New York, NY: Routledge.
Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation*. Lund, Sweden: Gleerup.
Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2012). *Using multivariate statistics*. New York, NY: Pearson.
Trausan-Matu, S., Stahl, G., & Sarmiento, J. (2007). Supporting polyphonic collaborative learning. *E-service Journal, Indiana University Press*, *6*, 58–74.
Thorndike, E. L., & Lorge, I. (1944). *The teacher's wordbook of 30,000 words*. New York, NY: Columbia University, Teachers College: Bureau of Publications.
Witten, I. A., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning and techniques*. San Francisco, CA: Elsevier.
Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.