B122164_Dissertation_10661-Words

anonymous marking enabled

Submission date: 14-Apr-2021 10:44AM (UTC+0100)
Submission ID: 149333446
File name: B122164_10661_words.pdf (2.42M)
Word count: 14906
Character count: 86773

The University of Edinburgh

School of Philosophy, Psychology and Language Sciences

Assessing New Entropy-Based Tools for Readability Assessment

B122164

BSc Cognitive Science

Supervisors: Hannah Rohde & Catherine Lai

Date of Submission: April 2021

Word count: 10,661

Abstract

Automated readability assessment has long been used by educators, publishers, and researchers to promote the accessibility of information for all audiences. Early formulae for predicting readability were first developed in the 1940s, using average word and sentence length to score each text with a readability grade. These are still widely used today, providing an indication of the education level required to successfully read and understand a given text. More recent methods for predicting reading difficulty make use of deeper linguistic features such as text cohesion and vocabulary difficulty. These newer methods may provide more accurate predictions of reading difficulty, however they are inaccessible to many writers and researchers due to the advanced natural language processing (NLP) tools and knowledge required to compute and interpret their output.

An alternative approach to readability assessment is the use of entropy to measure the 'surprisal' of a text. Surprisal is a key factor in reading difficulty, as readers form expectations about upcoming words in a text, and process predictable texts more easily than surprising ones. Entropy-based readability models were developed by Xing et al. (2008), who found that such models could out-perform traditional formulae when predicting text difficulty. These models have clear advantages over traditional readability formulae, and do not require advanced NLP techniques. This dissertation provides an external validation of these results in two domains: educational resources and manually simplified texts.

In the domain of educational resources, traditional formulae were found to be the best predictors of difficulty grade. This was contrary to the results reported by Xing et al. (2008), and suggests that their readability models do not generalise well to this test domain. However, in the domain of manually simplified texts, an entropy-based readability model significantly outperformed traditional formulae.

Acknowledgements

This dissertation would not have been possible without support and guidance from my supervisors: Hannah Rohde and Catherine Lai. I am very grateful for all their advice, and encouragement throughout the process – despite never getting to meet in person.

I would also like to thank all those at the University who have given me the skills and confidence to undertake this project, and fostered a genuine interest in research throughout my time as a student.

Thanks to Sarah, Pam and Kirsty for their lovely proofreading services and feedback.

Table of Contents

1. Introduction

The readability (or reading difficulty) of written texts is a measure of how easily they can be understood by readers. Naturally, some texts are more difficult to understand than others due to their subject matter, however many barriers to comprehension are rooted in linguistic complexity rather than information content. Therefore, a key factor in effective communication is to match the linguistic complexity of a text to the literacy level of readers, in order to improve understanding and engagement. For example, consider the following sentences:

1) *The cat sat on the mat.*
2) *On the mat sat the cat.*

The two sentences convey the same information to the reader, however sentence (2) is more difficult to read due to its unusual syntactic structure.

Writing at an appropriate reading difficulty is essential in education and public communication, as advocated by the Plain English Campaign (Plain English Campaign, n.d.). This is endorsed by both public and private sector bodies (Government Digital Service, 2016; The Plain Writing Act, 2010). By decreasing the reading difficulty of texts, writers improve the accessibility of information to all audiences; particularly those with low literacy levels such as second-language learners, children, and those with cognitive impairments.

In order to write and select written materials appropriate for their audiences, methods have been developed to label texts with a readability grade or score. These methods use either human judgements and responses, or automated measures based on the linguistic features of the text (Klare, 1974). Human judgements can provide subjective assessments of processing difficulty, which are considered to be more sensitive and reliable than automated measures (Schwartz et al., 1970). However, this method is not practical for assessing large volumes of text and is dependent on the reader's literacy level and domain knowledge as well as their subjective perception of cognitive effort. Automated measures of readability are objective in their assessments, using the linguistic features of a text as predictors of readability (Feng et al., 2010).

Traditional assessment formulae model the semantic difficulty of texts by measuring word length in terms of syllables, assuming that longer words are more difficult to read, or using a vocabulary list of words which are considered to be easily readable (Dale & Chall, 1948; Flesch, 1948; R. Gunning, 1952; McLaughlin, 1969). Sentence length is used to approximate syntactic complexity, assuming that longer sentences are more complex, without considering word order or syntactic structure. These assumptions are reflected in much of the general advice on improving readability: using short words and sentences makes information clearer to readers (Government Digital Service, 2016; IEEE, n.d.; The Plain Writing Act, 2010). Such assumptions may be true in the broad sense; however their key limitations are their disregard for word order, semantic content and the potential for longer words or sentences to improve readability (McNamara et al., 2010).

The extraction of more complex linguistic features from texts allows more cognitively motivated metrics to be used to model readability. These metrics capture lexical sophistication, syntactic complexity and text cohesion using state-of-the-art natural language processing (NLP) tools (Choi & Crossley, 2020). This approach addresses many of the limitations of traditional measures, capturing the effect of vocabulary and sentence structure on reading difficulty. The drawback of such models, however, is their complexity and the advanced NLP tools required to use them. Thus, their usefulness for research on readability and in writing and text simplification tools is limited.

An alternative measure to assess reading difficulty is the degree of predictability of text units, as readers process more predictable words and sentences more easily than unexpected ones (Jurafsky, 2003). This accounts for the differences in readability of sentences (1) and (2) (above), as the construction of sentence (1) makes it more predictable than sentence (2). Unpredictable information also contributes to the readability of texts, for example sentence (4) is less predictable than (3) due to its semantic meaning:

3) *The man bought the house.*
4) *The cat bought the house.*

In information theory, the predictability of a system can be characterised by its entropy (Shannon, 1948): systems with higher entropy are more unpredictable than those with low entropy. This allows the degree of predictability of a text to be quantified by its cross-entropy against a language model: how predictable the text is, given the reader's experience of language in general (Hale, 2006). The total entropy of a text characterises its degree of unpredictability, which corresponds loosely to its reading difficulty. However, a more sophisticated readability model should also consider semantic and syntactic difficulty. Such a model was first developed by Xing et al. (2008), who used a bigram language model, and average per-word and per-sentence entropy as proxies for semantic and syntactic difficulty. Two versions of this model were trained and validated on English textbook passages written for Chinese second-language English learners, modelling readability in different settings. The 'time-limited' model – designed to measure readability in contexts where the reader has limited time or resources to process a text – significantly outperformed traditional measures when grading passages by reading difficulty. This is a promising approach to automated readability assessment, as it overcomes some key weaknesses of traditional readability formulae without requiring advanced NLP techniques.

Despite their promising results, this entropy-based approach to readability assessment has not been validated in other contexts in which it may be useful. This dissertation thus implements the models described by Xing et al. (2008), and evaluates their validity for assessing readability in two domains: educational resources and manually simplified English texts. The results show that, in the domain of educational resources, the traditional Flesch-Kincaid formula is a better predictor of readability than either of the entropy-based models. In the domain of manually simplified texts, however, the 'time-limited' model significantly outperformed the Flesch-Kincaid formula. These results indicate that, in settings where information content is held constant, such as text simplification or certain research areas, an entropy-based readability model could provide a valuable tool to quantify and compare reading difficulty.

2. Literature Review

Readability is defined by the Oxford English Dictionary as "the ease with which a text may be scanned or read; the quality in a book, etc., of being easy to understand and enjoyable to read."(OED Online, 2020). Research on readability considers different variations on this definition, depending on the linguistic factors and audiences considered. The definition adopted for the purposes of this project is that of Dale and Chall (1949):

> The sum total (including the interactions of) all those elements within a given piece of printed material that affects the success that a group of readers have with it. That success is the extent to which they understand it, read it at an optimum speed, and find it interesting. (p. 23)

This is adapted for modern reading materials to include web-based text sources as well as printed materials. Poor readability, or high reading difficulty, means that readers will have little success in understanding and engaging with written materials. Improving readability is therefore of great interest to writers, publishers, and educators in order to communicate effectively with their audiences.

2.1 The Importance of Readability

The importance of selecting texts of appropriate reading difficulty is most obvious in education. Students of different ages vary greatly in their literacy levels and language exposure, and it is the task of educators to select reading materials appropriate for their students. The effect of readability on learning has been formally investigated, finding significant correlation between readability and recall (Rubenstein & Aborn, 1958), and lack of progress and frustration for students given reading materials that do not match their reading ability (T. Gunning, 2003).

The effect of readability on comprehension and information retention in adult readers is also well documented (Fass & Schumacher, 1978), as well as the tendency of readers to give up when confronted with a text that is too difficult for them (Dubay, 2004). This is particularly important in texts where it is essential that readers fully understand the information presented to them, such as in clinical and health settings (Ley & Florio, 1996; Worrall et al., 2020), and safety information (Sinyai et al., 2018; Yeomans, 2009). These studies show that when evaluating written materials using traditional readability measures, many materials are too difficult to read for the majority of the adult population. A famous case study of such a readability mismatch was presented by Wegner and Girasek (2003), who argued that many infant deaths due to traffic accidents could be attributed to the improper fitting of child safety seats, which could in turn be attributed to the reading difficulty of their installation instructions.

It could be argued that the difficulty in comprehension of technical information such as medical advice and safety instructions is due to a lack of motivation or domain knowledge in readers, however readability is also key concern when writing for specialist audiences. In financial reporting, the readability of reports is shown to have a significant impact on market responses (Kuang et al., 2020) and the behaviour of financial analysts (Lehavy et al., 2011), as poor readability results in greater uncertainty, leading to more negative responses and forecasts. The readability of software specification documents has also been cited as a key factor in protecting both clients and developers from misunderstandings in software requirements (Kanter et al., 2008).

One way to achieve low reading difficulty is for authors to write with readability in mind, ensuring that their material is accessible to all audiences. An alternative method is text simplification, defined as "the process of reducing the linguistic complexity of a text, while still retaining the original

information and meaning" (Siddharthan, 2014, p. 260). This can be performed manually or by automated procedures (Shardlow, 2014), with the aim of improving the accessibility of information by offering a simplified version of texts to match the ability of readers. Any process of text simplification requires an objective measure of reading difficulty in order to recognise high-difficulty passages that require simplification, and to assess the output of such simplification procedures (Aluisio et al., 2010; Vajjala & Meurers, 2014).

Readability assessment has thus been highlighted as a key consideration for all writers and publishers of written communications, as well as a vital element of any text simplification procedure.

2.2 How Readability is Measured

A comprehensive study of early readability assessment tools (Klare, 1974) divides assessment methods into three broad categories: guessing, testing, and predicting.

Educators and writers gain an intuition for estimating the readability of texts based on their experience and feedback from readers, and are usually able to match texts to their audience appropriately. The judgements of such skilled individuals was, in fact, used as the basis of early readability studies (Gray & Leary, 1935). This ability to 'guess' text readability relies on a wealth of experience with different reading materials and audiences and is thus not shared by many writers and researchers to whom readability assessment is of interest.

Methods to objectively measure readability include comprehension and reading-out-loud tests (Ekwall & Henry, 1968), and predictability measures such as the cloze procedure (Taylor, 1953). The cloze procedure involves substituting out words from the text and asking participants to guess the missing words: more predictable texts are considered to have lower reading difficulty. The validity of the cloze procedure as a measure of reading comprehension has been investigated in depth (Bormuth, 1968), showing significant correlation between predictability and comprehension difficulty and recommending that the cloze procedure be used for assessing the suitability of reading materials. This provides further evidence for the relationship between predictability and readability, on which the entropy-based measure of readability is based.

Procedures for measuring readability can provide useful evidence and guidance for teachers and writers, however conducting such tests is not a practical solution for large volumes of text. This motivates the use of predictive models such as readability formulae and the entropy-based measure implemented in this project.

Table 1 shows examples of 'traditional' readability formulae, dating from 1948 to 1975 (Begeny & Greene, 2014). These metrics are easily extracted from texts, using word and sentence lengths to model reading difficulty.

Table 1

Summary of Traditional Readability Formulae

Name(s)	Formula	Intended Use
Flesch Reading Ease (Flesch, 1948)	$Reading\ Ease = 206.835 - (0.846 * average\ syllables\ per\ word) - (1.015 * average\ words\ per\ sentence)$	First developed to improve the readability of newspapers. The resulting score ranges from 0 to 100, corresponding to the readability of a text.
Flesch-Kincaid, Automated Readability Index (Kincaid et al., 1975)	$Grade = (0.39 * average\ words\ per\ sentence) + (11.8 * average\ syllables\ per\ word) - 15.59$	Developed to grade US Navy training manuals. Adapts the Flesch Reading Ease formula to give a US Grade level, corresponding to the education level required to read the text.
Dale-Chall (Dale & Chall, 1948)	$Grade = (0.1579 * percent\ unfamiliar\ words) + (0.0496 * average\ words\ per\ sentence) + 3.6365$	Developed to score reading materials by suitability for students of different US grades. Familiar words were defined as those easily understood by 4th Grade students.
Gunning Fog Index (R. Gunning, 1952)	$Grade = (0.4 * average\ sentence\ length) + (percent\ words\ with\ more\ than\ two\ syllables)$	Developed for newspaper and textbook publishers. Grade represents the number of years of formal education required to read a text.
Simple Measure of Gobbledegook, SMOG (McLaughlin, 1969)	$Grade = 3 + \sqrt{number\ of\ words\ with\ more\ than\ two\ syllables}$	Three 10-word samples are taken from the text to compute its grade, corresponding to the level of education required to read a text.

Traditional measures of readability use sentence length as an approximation of syntactic difficulty. The assumption that longer sentences are more difficult to read may hold in the general sense, as they tend to have more complex structure, however it does not consider word order or syntactic construction. Word order has a significant impact on readability, as sentence structure can vary greatly between sentences containing the same lexical items:

5) *The mouse bit the cat that chased the dog that ran away.*
6) *The dog that the cat that the mouse bit chased ran away.*

This example, adopted from Lakretz et al. (2020), demonstrates the effect of word order on parsing difficulty, which cannot be detected by readability metrics based on sentence length. Furthermore, it has been found that sentence length does not affect reading fluency, when considered independently from syntactic complexity (Ratner & Sih, 1987). These limitations of sentence length as a metric to measure the syntactic difficulty of sentences are addressed by an entropy-based measure, as word order is used to calculate the average per-sentence entropy of the text.

Another key weakness of traditional readability formulae is their crude modelling of word difficulty. Most formulae approximate word difficulty by length (in terms of syllables), asserting that longer words are more difficult to read and understand. A more appropriate measure of word difficulty is frequency, as the use of more common words significantly improves reading comprehension (Marks et al., 1974). Due to computational limitations, early formulae could not assess word frequency, instead appealing to the correlation between word length and frequency (Sigurd et al., 2004). An

alternative measure of word difficulty is applied by the Dale-Chall formula (Dale & Chall, 1948), where a list of 3000 'familiar' words is used to differentiate between 'easy' and 'difficult' words. This is a limited approach as the familiar-words list reflects only the words that were most easily understood by a sample of students at the time that the list was compiled, which does not necessarily reflect the linguistic experience of readers today. The entropy-based readability measure uses word frequency directly in building the language model, providing a more accurate measure of word difficulty than traditional formulae.

In addition to their poor modelling of sentence and word difficulty, it must also be recognised that the data used to parameterise traditional readability formulae is now out of date. Many of these formulae were developed over 50 years ago for specific domain applications, using whatever training data were available at the time. This training data may not reflect the linguistic experience of readers today, undermining the validity of these formulae to grade texts for modern audiences.

The limitations of traditional readability formulae have been acknowledged (Hartley, 2016; Redish, 2000), motivating the development of more sensitive, cognitively-motivated measures. The Coh-metrix (Crossley et al., 2008; McNamara et al., 2010) was developed to capture vocabulary difficulty and text cohesion. Vocabulary difficulty is measured using word frequency. Cohesion is measured by sentence-sentence similarity (assessed by latent semantic analysis), use of causal connectives, and content-word overlap between sentences. Advanced NLP techniques - including analyses of lexical complexity, syntactic analysis and text cohesion - were also employed in the development of the Crowdsourced Algorithm for Readability (CAREC) (Crossley et al., 2019). These measures incorporate much deeper linguistic features than traditional formulae, addressing their weaknesses in measuring semantic and syntactic difficulty (Choi & Crossley, 2020). However, their use of advanced NLP techniques makes them inaccessible and opaque to many researchers and educators outside the field of linguistics.

Despite the widely acknowledged weaknesses of traditional readability formulae (Redish, 2000), researchers continue to employ them in readability studies. One active area of readability research is the accessibility of online health information such as allergy information (Tater, 2021), surgical patient advice (Besson et al., 2020), drug treatment guidance (Crawford-Manning et al., 2020) and public information about the ongoing COVID-19 pandemic (Worrall et al., 2020). Assessing and improving the readability of such resources is vital to improve public health and patient outcomes. However, these studies are all limited in validity by their exclusive use of traditional formulae: Flesch reading ease, Flesch-Kincaid, Gunning-Fog and SMOG. This reflects the preference of researchers for more easily interpretable readability measures, and their lack of access to more advanced methods.

The use of an entropy-based readability model aims to address some of the shortcomings of traditional readability formulae, while remaining accessible to users. The language model used to assess text entropy can take into account word order and frequency, and is open to adaptations which could better capture syntactic complexity. The model's principles and implementation do not require advanced NLP tools or domain knowledge, making it more useable in practical applications than other computational measures.

2.3 An Entropy-Based Approach to Readability

Entropy H is a measure of unpredictability of a variable with n possible outcomes (Shannon, 1948):

$$H = -\sum_{i=1}^{n} p_i \log(p_i) \quad (1)$$

Where each p_i is the probability of the variable taking value i

This corresponds to the difficulty of correctly guessing the variable's value when the probability distribution (each p_i) is known. A smaller entropy value means that it is easier to guess the variable's value, for example a loaded coin which always lands 'tails' has an entropy value of zero (as it is trivially easy to guess 'tails' every time). A fair coin has greater entropy (≈ 0.3), whereas a fair six-sided die has yet greater entropy (≈ 0.78), as the outcome is increasingly difficult to predict.

In these examples, the probability distribution over outcomes is known: if a coin is fair, then both 'heads' and 'tails' occur with 50% probability. In natural language processing, the probability distribution over linguistic units such as words, bigrams or sentences can be estimated from a training corpus. This corpus is assumed to be representative of the reader's experience of natural language.

During reading, readers use their experience of language to make predictions about what words are likely to come next in a sentence, processing more predictable words more easily than unexpected ones (Linzen & Jaeger, 2014; Venhuizen et al., 2019; Yan & Jaeger, 2020). Hence words with greater surprisal incur greater cognitive load and are inferred to convey more information than unsurprising words (Levy, 2013).

The predictability of a text can be quantified by the difference in probability distributions between the text and the language model: texts which diverge more from the reader's experience of language are more surprising and therefore harder to read. Cross-entropy $H(P, Q)$ is a measure of the difference between two probability distributions P and Q (Brownlee, 2019):

$$H(P, Q) = -\sum_{i=1}^{n} p_i \log(q_i) \quad (2)$$

Where p_i is the probability of the variable taking value i under P,
q_i is the probability of the variable taking value i under Q

The greater the divergence between two probability distributions, the greater their cross-entropy. If one of these distributions represents a person's expectations, and the other represents a new experience, then this cross-entropy represents how surprising the new experience will be, based on the person's expectations. For example, if a person is expecting a coin to land on 'heads' and 'tails' with equal probability, they will be more surprised to find a loaded coin which always lands 'heads' than one which lands 'heads' with 51% probability, as the former diverges more from their expectations.

Hence, the cross-entropy of a text against the language model corresponds to how much it diverges from the readers' experience of language in general, and therefore how difficult it is to read. When

referring to the entropy of a text, it should be assumed that this refers to the cross-entropy of the text against the language model.

The use of predictability to directly model readability is the basis of Xing et al.'s (2008) entropy-based measure. Word predictability is measured using a bigram language model, computing the conditional probability of seeing each word, given only the previous word in the sentence. The entropy of each word is thus the negative log of its conditional probability, and the sentence entropy is the sum of its constituent word entropies. For example, the entropy of the following sentence is computed as follows:

$$Sentence = The\ cat\ sat\ on\ the\ mat$$

$$Bigrams = \{(The\ cat), (cat\ sat), (sat\ on), (on\ the), (the\ mat)\}$$

Bigram probabilities:

$$P(The\ cat) = P(cat \mid The) = \frac{P(The\ cat)}{P(The)} \tag{3}$$

Where P(The cat) and P(The) are estimated from the training corpus

Bigram entropies:

$$H(The\ cat) = -\log\big(P(The\ cat)\big) \tag{4}$$

Where P(The cat) is computed by (3)

Sentence entropy:

$$H(The\ cat\ sat\ on\ the\ mat) = \sum_{b \in Bigrams} H(b) \tag{5}$$

Where each H(b) is computed by (4)

The total entropy of a text H_T is the sum of its constituent sentence entropies. This means that text entropy scales with text length as there are more terms in the summation, corresponding to the increasing difficulty of predicting the contents of longer texts. This total text entropy is used to compute average per-sentence entropy H_s and per-word entropy H_w, corresponding to sentence and word difficulty, respectively.

Xing et al.'s (2008) investigation found that all three metrics H_T, H_s and H_w had a linear relationship with the reading difficulty of textbook passages. This motivated their development of two linear models of readability:

1) The time-free model is a linear combination of H_s and H_w. This models readability with the assumption that there is no time pressure on the reader, so the total information contained in the text (H_T) should not affect readability. This ensures that the model does not simply scale readability with the length of a text.

2) The time-limited model is a linear combination of H_T, H_s and H_w. This models readability with the assumption that the reader has a limited amount of time or cognitive resources to process the text. Hence H_T should also be considered, as large amounts of information (longer texts) are harder to read in this setting.

Both models developed by Xing et al. (2008) were trained on passages in New Concept English textbooks (Alexander & Stoldt, 1972). This is a set of texts designed for Chinese second-language English (L2) learners, which are graded by reading difficulty. The models were then tested on graded passages from the College English Textbook (Skwire & Wiener, 1998), another English textbook written for Chinese L2 students. Readability scores from both models significantly correlated with difficulty grade in the test data, while the time-limited model outperformed both the time-free model and the traditional Flesch-Kincaid formula.

The results presented by Xing et al. (2008) indicate that a simple entropy-based linear model can accurately predict reading difficulty, however both their training and test data were from the limited domain of L2 English textbooks. Based on this evidence, the validity of their model in other domains is unclear.

2.4 Study Goals

The purpose of this dissertation is to assess the validity of the entropy-based measures developed by Xing et al. (2008), compared to traditional readability formulae. The entropy-based measures, and the Flesch-Kincaid readability formula will be used to score the readability of graded educational resources and manually simplified texts, to evaluate the strength of association between readability scores and labelled text difficulty. Each readability score will then be used as a predictor of reading difficulty, to compare the predictive validity of each measure.

Educational resources are a useful domain for evaluating the entropy-based readability models, as they are already graded by the publishers to indicate the difficulty level of content. If the reading difficulty scores generated by the models are accurate predictors of graded difficulty, the models may provide a useful tool for educators and publishers to assess the difficulty of new resources.

Manually simplified texts are another relevant domain in readability research, as quantifying the difference between difficulty levels is vital for developing and evaluating text simplification tools. These test texts provide matched samples, where different versions of the same articles have been written at different reading difficulty levels. If the entropy-based readability models can accurately distinguish between these difficulty levels, they may be used in the development and evaluation of text simplification tools to quantify the difference in reading difficulty between the input and output of such tools.

The entropy-based models, if validated in these domains, could provide useful readability measures to assist educators, publishers, and the developers of text simplification tools.

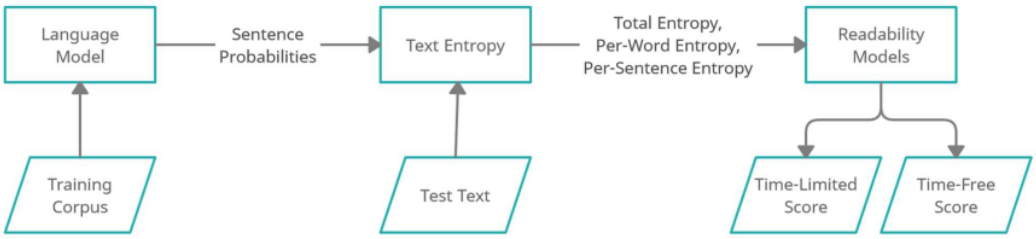
3. Methods

In this study, the entropy-based readability models developed by Xing et al. (2008) were implemented and assessed on four sets of graded texts from different sources. The implementation is intended to match that of Xing et al. as closely as possible, to independently assess the performance on their model on alternative test data. The model was built and run in Python, using the NLP functionality provided by the Natural Language Toolkit (NLTK) (Bird & Loper, 2020).

Figure 1 illustrates the pipeline for computing entropy-based readability scores for each test text, using the language model built from the training corpus.

Figure 1

Overview of Components of the Readability Models



For each test text, the time-limited and time-free model scores can be computed using the language model and readability models. The Flesh-Kincaid readability grade for each text is also computed, to compare the performance of the entropy-based models to a traditional readability formula. All three metrics represent reading difficulty (as opposed to reading ease), so more difficult texts will have higher scores.

Four corpora of test texts were used, each providing a set of text passages labelled with their reading difficulty. In each corpus, the model scores were tested for a statistically significant relationship with labelled difficulty grade. A predictive model was then fitted to evaluate how well each readability measure predicts graded reading difficulty. This allows models to be compared for their predictive accuracy in each domain.

3.1 The Language Model

A language model is a probability distribution over sequences of words, estimated from a training corpus. Each word in a sentence is assigned a probability based on the words that have already been seen, and the probability of the sentence is the product of its constituent word probabilities. A bigram language model makes the simplifying assumption that the probability of seeing each word is dependent only on the previous word, so the bigram probability of seeing a sequence of words w_1 to w_n is given by:

$$P_{Bigram}(w_1, \dots, w_n) = \prod_{i=2}^{n} p(w_i \mid w_{i-1}) \tag{6}$$

Where $p(w_i|w_{i-1})$ is the conditional probability of seeing word w_i given the previous word w_{i-1}

Each of the conditional probabilities $p(w_2|w_1)$ are estimated from the training corpus using bigram frequency:

$$p(w_2|w_1) = \frac{p(w_1, w_2)}{p(w_1)} \approx \frac{f(w_1, w_2)}{f(w_1)} \quad (7)$$

Where $f(w_1, w_2)$ is the frequency of the bigram (w_1, w_2) in the training corpus, and $f(w_1)$ is the frequency of w_1 in the training corpus

Sparsity is often a problem when using this estimation method, as any bigrams that do not appear in the training corpus will be assigned zero probability, which will in turn cause the entire sentence in which they appear to be assigned zero probability.

In order to assign nonzero probability to unseen bigrams, Witten-Bell Smoothing was applied to the bigram probability distribution. This smoothing method allows unseen bigrams (w_1, w_2) to be assigned a nonzero probability based on the diversity of words which follow w_1 in the training corpus, rather than alternative smoothing methods which uniformly distribute probability mass to unseen events (Sokolov, 2015). For example, the word *spite* is followed by seven different words in the training data, whereas *many* is followed by 1353 different words. This means that the probability of seeing a new bigram which starts with *many* should be greater than the probability of seeing a new bigram starting with *spite*. Witten-Bell smoothing assigns probabilities to unseen bigrams to reflect this 'diversity of histories': in this case the smoothed probability assigned to the bigram "*many tweets*" is eight times greater than the smoothed bigram probability for "*spite tweets*", despite neither of these bigrams appearing in the training corpus. This reflects how surprised the reader would be to see a new word following the previous one, given their linguistic experience.

Single-word sentences do not contain any bigrams, so the sentence probability is computed by a unigram language model estimated from word frequencies in the training corpus:

$$P_{Unigram}(w) \approx \frac{f(w)}{Number\ of\ words\ in\ the\ training\ corpus} \quad (8)$$

Lidstone smoothing was applied to the unigram probability distribution to assign nonzero probability to unseen words. This uniformly distributes a small amount of probability mass from words in the training corpus to unseen words.

The bigram language model uses word order and frequency to estimate sentence probability, however it fails to take into account the wider context in which a word occurs. Higher order n-gram models compute word probabilities conditioned on a longer preceding sequence rather than just the preceding word, however they suffer from a greater degree of sparsity and require a much larger training corpus to give good probability estimates (Jurafsky & Martin, 2020).

An alternative language model is a probabilistic context-free grammar (PCFG), which uses grammatical production rules and their associated production probabilities to parse each sentence (Hammond, 2006). These production rules can be used in a language model to estimate the probability of each word in the sentence, based on the parse tree constructed over the words that have been seen so far (Linzen & Jaeger, 2014). This method may provide more accurate estimates for the probability of seeing each word in the sentence, as it considers both vocabulary and syntactic structure over all preceding words. A PCFG language model requires a large, parsed training corpus, and greater computational resources when evaluating sentence probabilities than an n-gram model. This represents a potential development for the entropy-based readability model, however the

availability of training data makes the use of syntax-free bigram models more practically useful at present.

3.2 The Training Corpus

The training corpus used in this study is the same as that used by Xing et al. (2008), so that their model can be reproduced and independently validated. This training corpus consists of:

- The Brown Corpus of American English, compiled from texts published in 1961 (Francis & Kucera, 1979).
- The Lancaster-Oslo-Bergen (LOB) Corpus of British English, compiled as the British counterpart to the Brown corpus of texts published in 1961 (Johansson et al., 1986)
- The Freiburg-Brown Corpus of American English; an updated edition of the Brown Corpus using texts published in 1991 (Hundt et al., 1999).
- The Freiburg-LOB Corpus of British English; an updated edition of the LOB corpus using texts published in 1991 (Hundt et al., 1998).

The total size of the training corpus is just under 5 million words (4,760,957), covering English texts from a variety of genres (see Appendix A). This is designed to be representative of the linguistic experience of a typical reader, so the resulting language model reflects the likelihood of encountering any given word sequence in a 'typical' English text.

3.3 Computing Text Entropy

The entropy of a text is equal to its negative-log probability (Equation 1). For each sentence in the text, the probability of its constituent bigrams or unigram (in the case of single-word sentences) are given by the smoothed language model. These probabilities are then converted to entropies by taking their negative logarithm, and the product over probabilities in each multi-word sentence (Equation 6) becomes a sum over entropies:

$$H_{multi-word}(w_1, \dots, w_n) = \sum_{i=2}^{n} -\log_2 p(w_i|w_{i-1}) \tag{9}$$

Where n > 1 and $p(w_i|w_{i-1})$ is estimated from the language model

For single-word sentences, there is only a single term:

$$H_{single-word}(w) = -\log_2 p(w) \tag{10}$$

Where p(w) is estimated from the language model

The total entropy H_T of a text is computed as the sum over its constituent sentence entropies:

$$H_T = \sum_{for\ S \in multi-word\ sentences} H_{multi-word}(S) + \sum_{for\ S \in single-word\ sentences} H_{single-word}(S) \tag{11}$$

From this total entropy, the average per-word entropy H_w, and per-sentence entropy H_s are computed as follows:

$$H_w = \frac{H_T}{Number\ of\ words\ in\ the\ text} \tag{12}$$

$$H_s = \frac{H_T}{Number\ of\ sentences\ in\ the\ text} \quad (13)$$

Texts units with greater entropy are considered to be more difficult to read, as they contain more unpredictable bigrams. Computing the average per-word and per-sentence entropy provides an estimate for the predictability, and therefore readability, of each word and sentence in the text.

3.4 The Readability Models

The readability models developed by Xing et al. (2008) use linear combinations of H_T, H_s and H_w to measure reading difficulty. The total text entropy H_T represents the total amount of information in the text, as more surprising (higher entropy) texts are considered to convey more information to the reader. Thus H_w and H_s represent the average amount of information conveyed by each word and sentence, respectively. These in turn correspond to the semantic and syntactic difficulty of the text, as words and sentences which convey more information are assumed to be more difficult to read.

The time-free model is a linear combination of semantic and syntactic difficulty, designed to measure readability where there is no time-pressure on the reader. The total amount of information contained in the text (H_T) does not contribute to this model, so the readability score will not scale with text length. The model coefficients were estimated by Xing et al. using maximum-likelihood estimation on passages from the New Concept English textbooks (Alexander & Stoldt, 1972). The resulting linear model computes the reading difficulty of each text as follows:

$$Time\ Free\ Score = -57.445 + 15.199\,H_w + 1.932\,H_s \quad (14)$$

The time-limited model is a linear combination of all three metrics H_T, H_s and H_w. The total text entropy does contribute to reading difficulty in this case, as it is assumed that the reader has limited time or cognitive resources to process the text. This means that texts containing more information should be scored with a greater reading difficulty. Model coefficients were estimated using the same method and training data as the time-free model, giving the following linear model:

$$Time\ Limited\ Score = -8.093 + 5.023\,H_w + 1.245\,H_s + 0.055\,H_T \quad (15)$$

In order to compare the entropy-based readability models to traditional measures, the Flesch-Kincaid readability grade for each text was also calculated (Kincaid et al., 1975). The implementation for this readability formula was provided by the Python Readability Metrics package (DiMascio, 2019).

3.4 The Test Data

Four sets of test data were used to evaluate the utility of each readability measure for predicting labelled reading difficulty in different domains. In all four test corpora, texts are labelled with a difficulty level.

In all test corpora, a positive correlation is expected between labelled difficulty and the three readability measures. The results reported by Xing et al. (2008) suggest that, in the domain of educational resources, the time-free model may outperform traditional measures. In the domain of manually simplified texts, the readability measures should detect significant differences in reading difficulty between grades, quantifying the effect of manual simplification.

3.4.1 Textbook Passages

A corpus of school textbooks was used to investigate whether the models' readability scores differ significantly between different school grades. This corpus consists of science and social science

textbooks written for American middle-school grades 6, 7 and 8 (see Appendix B). The reading difficulty of textbooks is expected to increase with school grade, however there is also an increase in conceptual difficulty so the true relationship between grade and readability is likely to be nonlinear.

3.4.2 Graded Reading Materials

While textbook passages are assumed to increase in difficulty with grade, English for Speakers of Other Languages (ESOL, n.d.) provides a set of graded text passages specifically designed to increase in reading difficulty. A corpus of 42 articles was compiled from the ESOL reading resources (ESOL, n.d.), each graded with a difficulty grade between 5 and 13; it is important to note, however, that these reading difficulty grades do not refer to the middle-school grades used in the textbook corpus.

3.4.3 OneStop Corpus of Manually Simplified Texts

The OneStop Corpus was compiled for use in automatic readability assessment and text simplification tools (Vajjala & Lučić, 2018). It consists of 189 texts, each written at three difficulty levels: elementary, intermediate, and advanced.

The text written at each difficulty level is intended to convey the same information, so the difference in readability scores between the different versions should be entirely due to linguistic factors rather than any difference in information content.

3.4.4 Wikipedia and SimpleWiki Articles

SimpleWiki is a parallel version of Wikipedia, with simplified articles written to convey the same information as the original page in a more readable format (Wikipedia, n.d.-b). The top 100 most-viewed Wikipedia pages were selected for analysis, excluding lists and pages logging many automated or unintentional views (see Appendix C). For each of these pages, the text contents of the English Wikipedia and Simple English Wikipedia were extracted, and readability scores for each were computed.

4. Results

This section describes the statistical methods used to analyse each test corpus, and the predictive accuracy of each readability measure for that corpus. The distribution of text lengths and readability scores is examined across different difficulty levels, as well as the relationship between difficulty level and readability scores. The number of labelled difficulty levels differs between corpora, necessitating different analytical approaches, such as group comparisons between different grades (where there are fewer distinct grades), or the correlation between grade and difficulty score (where there are more distinct grades).

In all four test corpora, an appropriate predictive model was fitted for each readability measure, and the cross-validated predictive accuracy computed to rank the readability measures as predictors of difficulty grade.

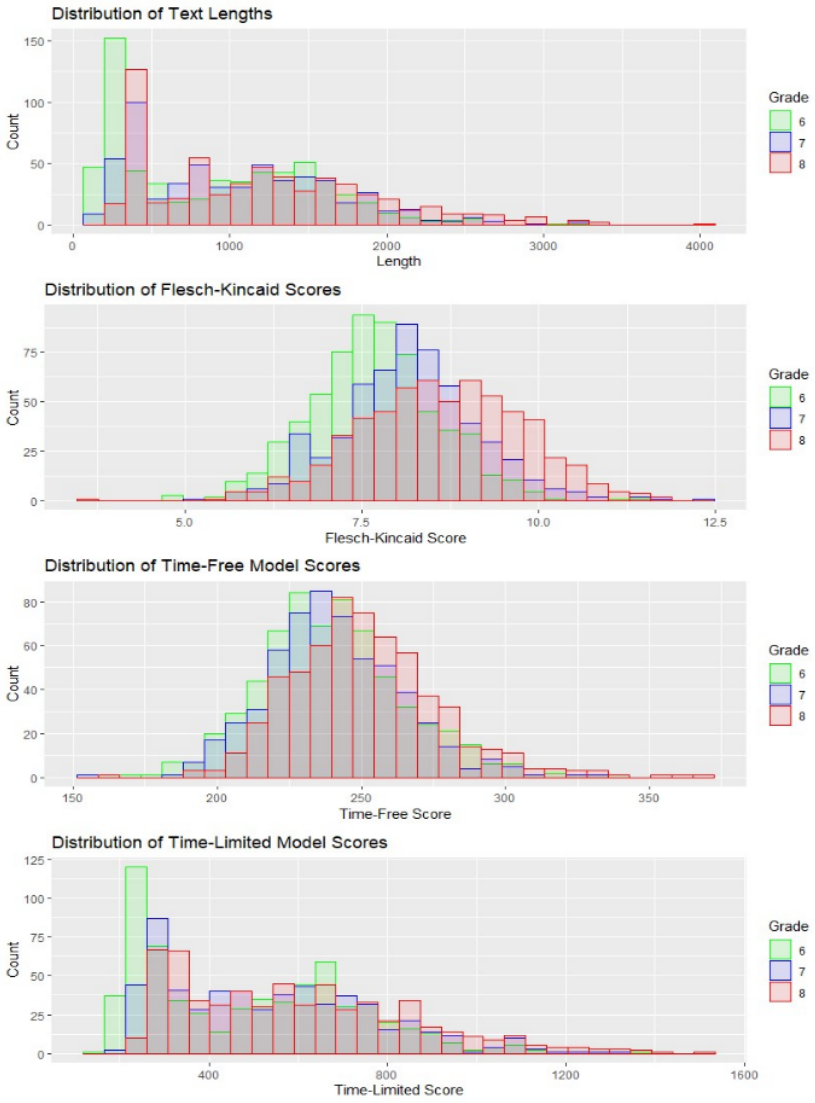
4.1 Textbook Passages

Each textbook is divided into units, chapters, and sections (N=1810), where each section is, on average, 1000 words in length (μ=1045, σ=661). Textbook sections were labelled with the school-grade of their intended audience, and readability scores for each section were computed.

Figure 2 shows the distribution of text lengths and model scores in each textbook grade. Both the Flesch-Kincaid and Time-Free model scores are normally distributed in each grade, with an increase in reading difficulty at higher grades. The distribution of Time-Limited model scores appears to mirror the distribution of text lengths, suggesting that this measure is heavily dependent on text length.

Figure 2

Distribution of Text Lengths and Readability Scores of Textbook Sections



The variance in Flesch-Kincaid score differs significantly between grades, failing Levene's test for equal variances. Welch's one-way ANOVA was performed to compare group means without equal variances, finding significant differences between grades for all three readability measures (see Appendix D).

Table 2 shows pairwise comparisons between grades, reporting Tukey's Honest Significant Difference between readability scores in each grade, as well as the 95% confidence interval for the difference in means.

Table 2

Tukey's Honest Significant Difference in Means

Measure	Grades	Difference in Means	95% Confidence Interval
Flesch-Kincaid	7-6	0.47**	[0.33, 0.61]
	8-6	0.94**	[0.80, 1.08]
	8-7	0.48**	[0.33, 0.62]
Time-Free	7-6	0.56	[-2.79 3.91]
	8-6	11.44**	[8,13, 14.76]
	8-7	10.88**	[7.48, 14.28]
Time-Limited	7-6	47.68**	[15.04, 80.32]
	8-6	110.06**	[77.77, 142.36]
	8-7	62.38**	[29.30, 95.46]

**p<.01 (two-tailed)

All three readability measures found the greatest difference in means between grades 6 and 8, as expected as this is the largest difference in education level of their intended audiences. A larger increase in readability score is found between grades 7 and 8 than between 6 and 7 for all three readability measures. These results suggest a monotonous, nonlinear relationship between reading difficulty and textbook grade.

The strength of association between each readability measure and textbook grade can therefore be assessed by their rank-correlation. The rank-correlation between each measure and textbook grade, and correlations between different readability measures are reported in Table 3.

Table 3

Spearman's Rank Correlation Coefficients in the Textbook Corpus

Measure	Textbook Grade	Flesch-Kincaid	Time-Free	Time-Limited
Textbook Grade	1			
Flesch-Kincaid	0.36**	1		
Time-Free	0.18**	0.37**	1	
Time-Limited	0.20**	0.14**	0.22**	1

**p<.01

There is a significant positive correlation between all readability measures and textbook grade, with Flesch-Kincaid grade correlating most strongly. There are also significant positive correlations among all three readability measures.

For each readability measure, an ordinal logistic regression classifier was fitted to predict textbook grade using the model score as a predictor. Ordinal logistic regression is an extension of binary logistic regression, using the ordering of outcomes (grades) to fit a coefficient for the predictor (model score). This coefficient is used to compute the odds ratio of being in a higher textbook grade, given a unit increase in readability score. For example, an odds ratio of 1.5 means that a unit increase in readability score makes a text 50% more likely to belong to a higher-grade textbook, and an odds ratio of 1 means that readability score has no effect on textbook grade. It is important to note that the scales of the different readability scores vary greatly, as Flesch-Kincaid score is designed to represent a school-grade between 1 and 13, whereas the entropy-based models operate on a scale designed to range into the hundreds.

For each classifier, 10-fold cross-validation was performed to assess the predictive accuracy of the model on held-out data. In turn, each fold is treated as a held-out test dataset, while the classifier is trained on the remaining data. The classifier is then used to predict the grade of each text in the test data. The accuracy and mean absolute error (MAE) of these predictions, given by Equations 16 and

17, were computed and averaged over all cross-validation folds. The mean accuracy of each model indicates the proportion of texts which were correctly classified, and MAE indicates the average discrepancy between predicted and actual grades. These results are reported in Table 4.

$$Accuracy = \frac{\#\ correct\ predictions}{\#\ predictions} \tag{16}$$

$$Mean\ Absolute\ Error = \frac{1}{\#\ predictions} \sum_{t\ \in test\ data} \|\hat{g}_t - g_t\| \tag{17}$$

Where $\hat{g}_t$ is the predicted grade of text t, and g_t is its actual grade

Table 4

Ordinal Regression Models fitted for the Textbook Corpus

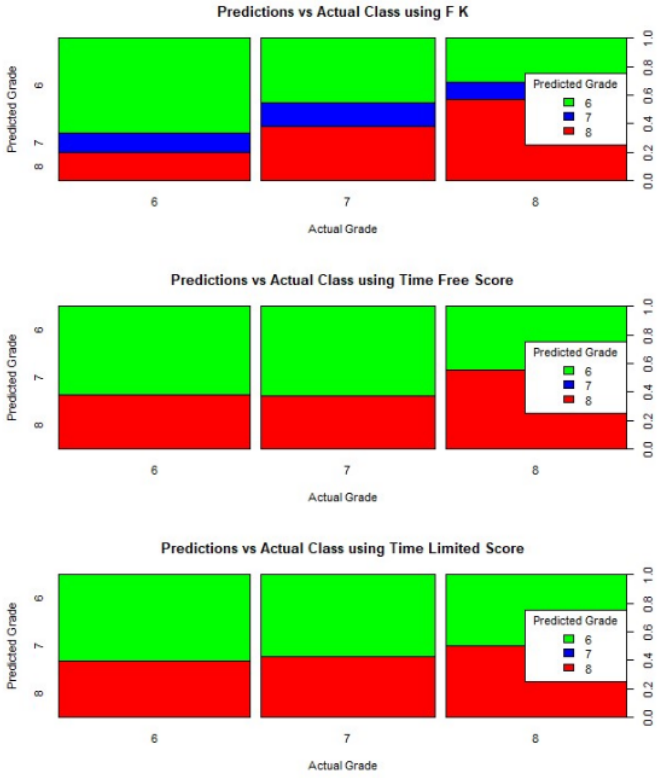
Measure	Odds Ratio	Mean Accuracy	Mean MAE
Flesch-Kincaid	1.932	48%	0.70
Time-Free	1.014	40%	0.88
Time-Limited	1.001	38%	0.92

The odds ratios in Table 4 are all greater than one, reflecting the positive correlation between each readability measure and textbook grade. This means that for all three measures, an increase in readability score corresponds to an increased probability of the text being in a higher grade. The different scales on which each of the measures operates makes a comparison between odds ratios meaningless, however the mean accuracy and mean absolute error in cross-validation can be used to rank the models by their predictive performance. The Flesch-Kincaid readability grade is shown to be the best predictor of textbook grade, while the two entropy-based measures showed similar performance. Contrary to the results of Xing et al. (2008), the time-free model performed better than the time-limited model in this domain.

Model predictions were also compared to the actual grade of each text, when training and predicting on the entire Textbook corpus. These predictions are plotted in Figure 3, indicating the proportion of texts in each grade that were correctly classified by each model.

Figure 3

Predicted vs Actual Grades for the Textbook Corpus



This illustrates the nature of classification errors, showing that grade 7 texts are most likely to be misclassified by all three models. This is consistent with the distributions in Figure 2, where the distribution of scores for grade 7 texts are most difficult to distinguish from the other grades.

4.2 Graded Reading Materials

Each article (*N*=42) from the ESOL site (ESOL, n.d.) was downloaded and manually edited to exclude any web-markup, headings or image captions. The remaining text component was annotated with its labelled difficulty grade, and readability scores were computed.

Figure 4

Relationship Between Text Length, Readability Score, and Reading Difficulty of ESOL Passages

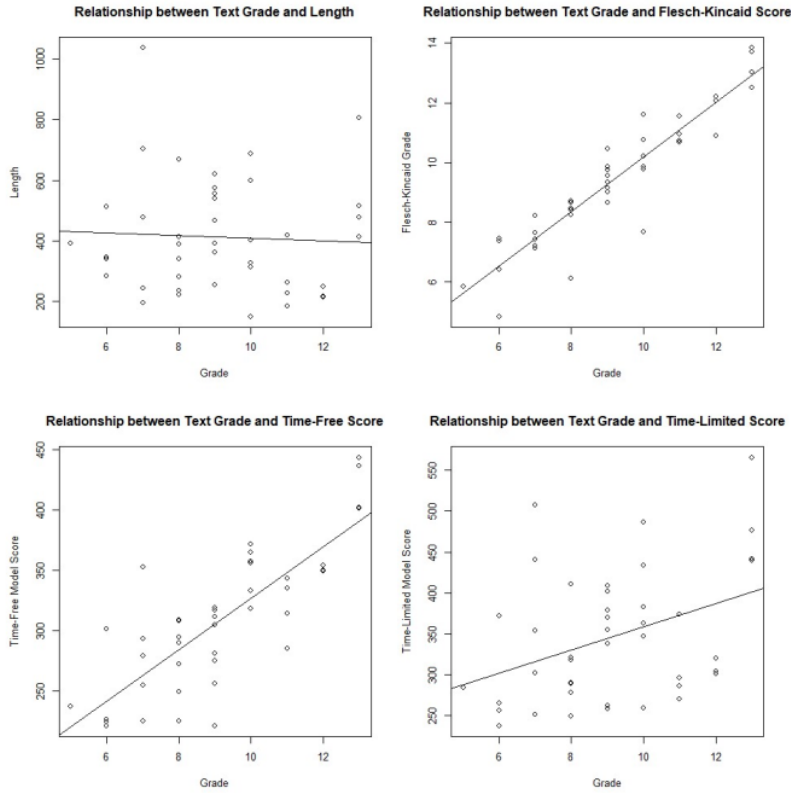


Figure 4 shows the relationship between text length and readability scores, and difficulty grade. All three readability measures display a positive linear relationship with difficulty grade. The strength of this relationship is assessed by Spearman's rank correlation test, which does not assume that the variables are normally distributed. The rank-correlation between readability measures is reported in Table 5.

Table 5

Spearman's Rank Correlation Coefficients in the ESOL Corpus

Measure	Difficulty Grade	Flesch-Kincaid	Time-Free	Time-Limited
Difficulty Grade	1			
Flesch-Kincaid	0.94**	1		
Time-Free	0.77**	0.76**	1	
Time-Limited	0.36*	0.45**	0.51**	1

*p<.05, **p<.01

The strongest correlation is found between Flesch-Kincaid readability score and difficulty grade, and there are significant positive correlations among all three readability measures. For each measure, a linear regression model was fitted to predict the difficulty grade of texts, using the readability score as a predictor. The coefficient generated by this model indicates the expected increase in difficulty grade, given a unit increase in readability score. 10-fold cross-validation was performed to evaluate the predictive accuracy of each model, and the mean MAE (Equation 17) across all cross-validation folds is reported in Table 6.

Table 6

Linear Models Fitted for the ESOL Corpus

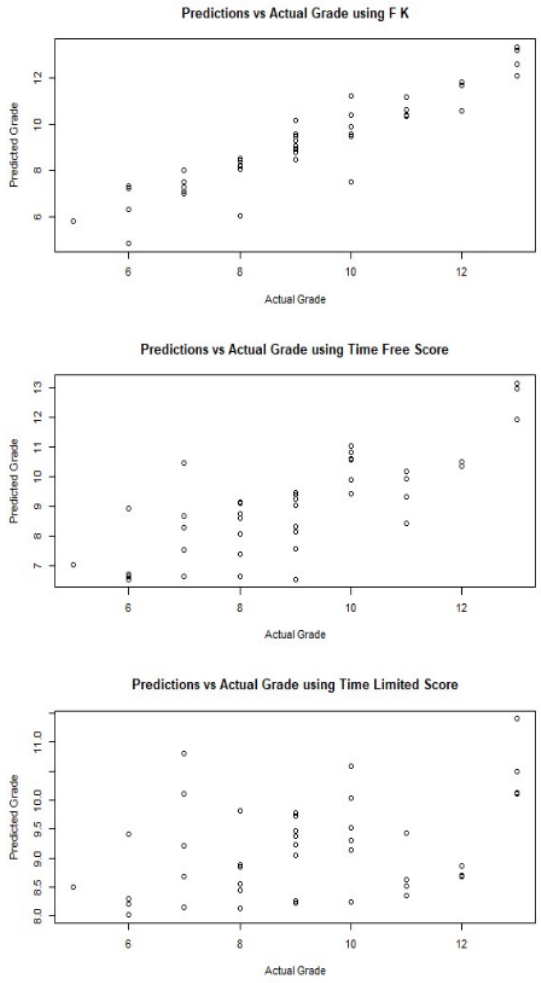
Measure	*Coefficient*	*Mean MAE*
Flesch-Kincaid	0.94	0.63
Time-Free	0.03	1.10
Time-Limited	0.01	1.66

The positive coefficients for all three models reflect the positive correlation between each readability score and difficulty grade. The mean MAE of each model can be used to rank them by their predictive accuracy, as a smaller MAE indicates that predictions are, on average, closer to the true grade. The results in this domain are consistent with those observed for the textbook corpus: the best predictor of labelled difficulty grade is the Flesch-Kincaid score, while the time-free model gives more accurate predictions than the time-limited model.

Model predictions were also compared to the actual grade of each text, when training and predicting on the entire corpus. Figure 5 illustrates the relationship between model predictions and actual labelled difficulty grade. If all predictions were correct, this figure would show a perfect linear relationship.

Figure 5

Predictions vs Actual Class for the Wikipedia/SimpleWiki Corpus



The plots in Figure 5 are consistent with the mean absolute errors for these models (Table 6): predictions from the Flesch-Kincaid model most closely match the labelled difficulty of texts.

4.3 OneStop Corpus of Manually Simplified Texts

The texts in this corpus are written at three difficulty levels: Elementary, Intermediate and Advanced. These are coded as factors with levels 1, 2 and 3, respectively.

Figure 6

Distribution of Text Lengths and Readability Scores of OneStop Articles

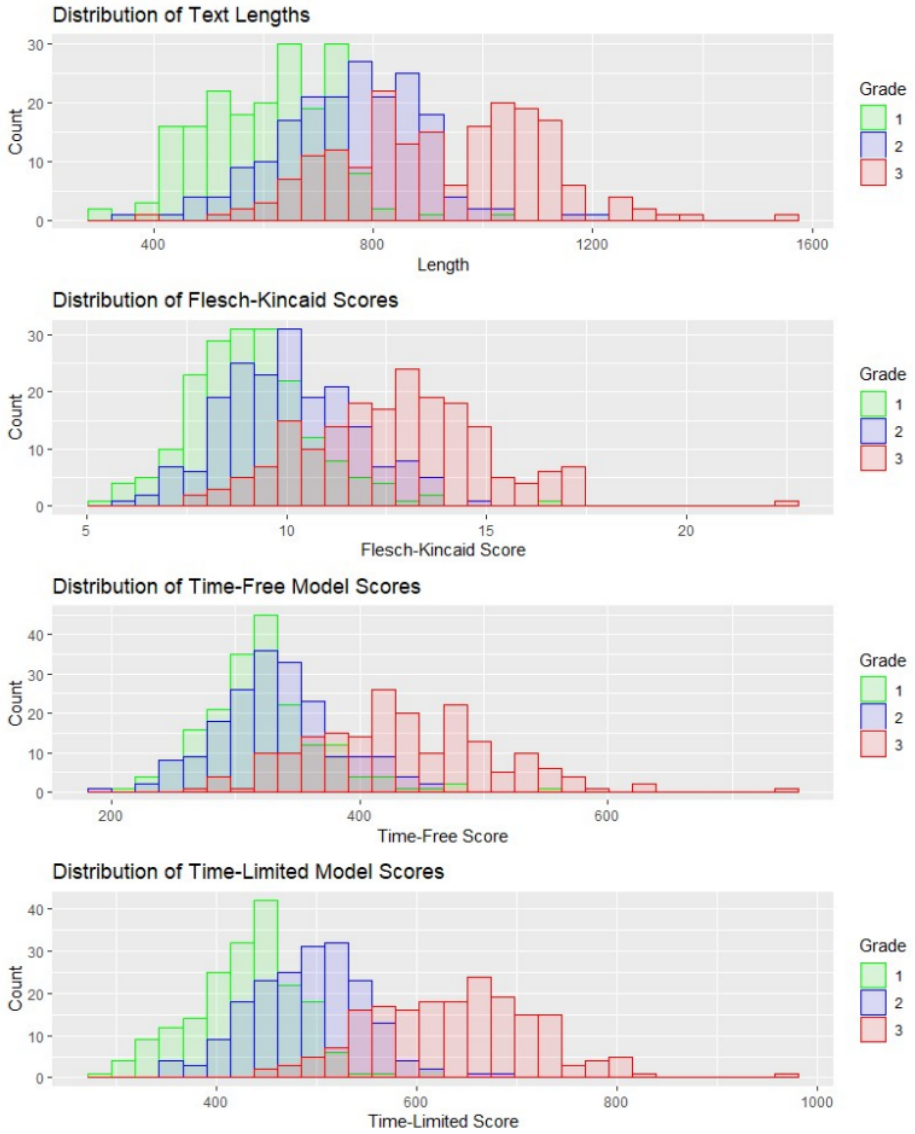


Figure 6 illustrates the distribution of text lengths and readability scores in each difficulty grade. All three readability scores are approximately normally distributed, showing an increase in readability score with difficulty grade. The variance in readability scores differs significantly between grades, failing Levene's test for equal variances. This violates the assumptions required for a repeated-

measures ANOVA, so the non-parametric Friedman test was used to detect significant differences in readability scores between groups. Significant differences were found between difficulty grades for all readability measures (see Appendix D).

Multiple pairwise t-tests were used to measure the difference in readability scores between difficulty levels. The mean difference in readability scores between each pair of difficulty levels is reported in Table 7.

Table 7

Paired-Samples T-tests for Significant Differences Between OneStop Difficulty Grades

Measure	Grades	Mean Difference	95% Confidence Interval
Flesch-Kincaid	2-1	0.93**	[0.76, 1.11]
	3-1	3.61**	[3.38, 3.84]
	3-2	2.68**	[2.47, 2.89]
Time-Free	2-1	10.46**	[4.62, 16.33]
	3-1	112.20**	[104.30, 120.10]
	3-2	101.73**	[94.30, 109.16]
Time-Limited	2-1	59.99**	[53.86, 66.12]
	3-1	207.39**	[197.50, 217.27]
	3-2	147.40**	[139.53, 155.27]

**p<.01

As expected, all three measures found significant increases in readability score as difficulty grade increases. There is a much larger increase in readability scores between grades 2 and 3, than between grades 1 and 2. This suggests a monotonous, nonlinear relationship between reading difficulty and grade.

The strength of association between each readability measure and textbook grade can therefore be assessed by their rank-correlation. The correlation between different readability measures is reported in Table 8.

Table 8

Spearman's Rank Correlation Coefficients in the OneStop Corpus

Measure	Difficulty Grade	Flesch-Kincaid	Time-Free	Time-Limited
Difficulty Grade	1			
Flesch-Kincaid	0.62**	1		
Time-Free	0.60**	0.89**	1	
Time-Limited	0.80**	0.65**	0.74**	1

**p<.01

For each readability measure, an ordinal regression classifier was fitted to predict the difficulty grade of texts based on each readability score. The odds ratio, mean cross-validated accuracy and mean MAE for each model are reported in Table 9.

Table 9

Ordinal Regression Models fitted for the OneStop Corpus

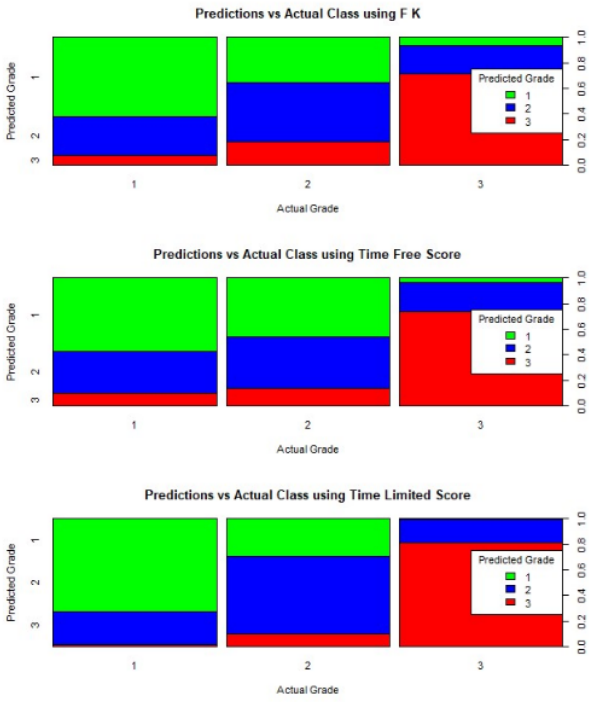
Measure	Odds Ratio	Mean Accuracy	Mean MAE
Flesch-Kincaid	2.016	59%	0.45
Time-Free	1.022	58%	0.46
Time-Limited	1.029	71%	0.29

In this domain, the time-limited model is the best predictor of difficulty grade, while the performance of the Time Free model and Flesch Kincaid as predictors are similar.

Model predictions were also compared to the actual grade of each text, when training and predicting on the entire OneStop corpus. Figure 7 illustrates the predicted grades of texts in each labelled difficulty grade.

Figure 7

Predicted vs Actual Grades for the OneStop Corpus



All three models correctly classify the majority of Advanced (grade 3) texts, while both the Flesch-Kincaid and Time Free models are most likely to misclassify Intermediate grade texts.

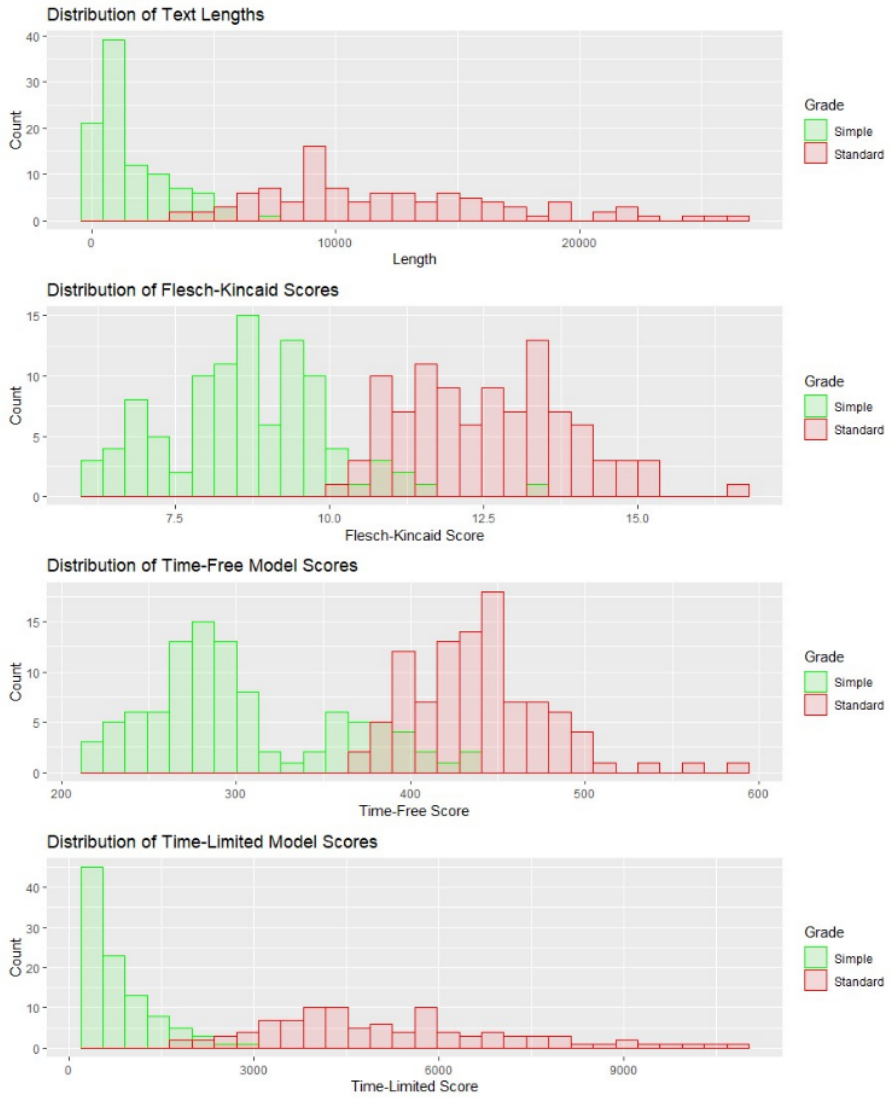
4.4 Wikipedia and SimpleWiki Articles

Lastly, the Wikipedia corpus is analysed. In this corpus, the largest difference in reading difficulty is expected between grades as there are only two labelled difficulty levels, written for vastly different audiences. This makes the prediction problem easier, so the best predictive accuracy is expected in this test corpus.

The text contents of each Wikipedia and SimpleWiki article was extracted, excluding any web-markup and image captions. Figure 8 shows the distribution of text lengths and readability scores at each difficulty level.

Figure 8

Distribution of Text Lengths and Readability Scores of Wikipedia and SimpleWIki Articles



All three readability measures show a clear distinction between simplified and standard articles. It is also clear in Figure 8 that the simplified articles are much shorter than standard articles.

The mean differences in readability scores between standard and simplified articles were assessed using paired-samples t-tests. The mean difference in readability scores between Standard and Simple articles is reported in Table 10.

Table 10

Paired-Samples T-tests for Significant Differences between Standard and Simple Wikipedia Articles

Measure	Mean Difference	95% Confidence Interval
Flesch-Kincaid	3.92**	[3.60, 4.23]
Time-Free	136.26**	[124.88, 147.64]
Time-Limited	4418.85**	[4042.82, 4794.89]

**p<.01

As expected, all three measures found that the readability scores of standard articles are significantly higher than simplified articles. Logistic regression models were fitted, using each readability score in turn to predict the difficulty level of texts. The odds ratio generated by each model represents the odds of an article being classified as 'Standard' rather than 'Simple', given a unit increase in readability score. The mean accuracy and mean MAE over 10-fold cross validation are reported in Table 11.

Table 11

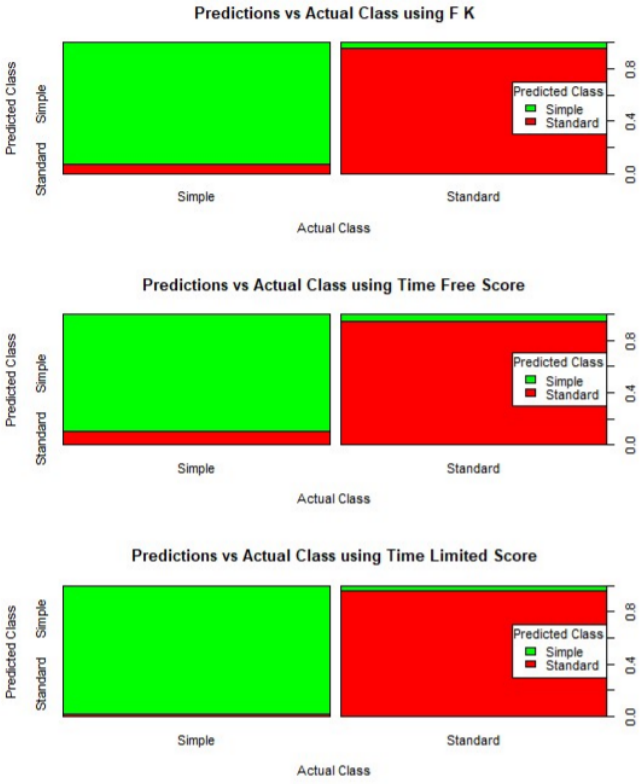
Logistic Regression Models to Classify Standard and Simple Wikipedia Articles

Measure	Odds Ratio	Mean Accuracy	Mean MAE
Flesch-Kincaid	12.99	94%	0.06
Time-Free	1.07	92%	0.08
Time-Limited	1.01	97%	0.03

Model predictions were also compared to the actual grade of each text, when training and predicting on the entire corpus. The proportion of correctly classified texts in each grade is illustrated in Figure 9.

Figure 9

Predictions vs Actual Class for the Wikipedia/SimpleWiki Corpus



The plots in Figure 9 are consistent with the predicted accuracy of each model (Table 11), as all three models correctly classified the vast majority of articles. The time-limited model was the best predictor of difficulty level in this domain, consistent with results in the OneStop corpus.

5. Discussion

The purpose of this investigation is to investigate the validity of entropy-based readability measures (Xing et al., 2008) for predicting the labelled difficulty of texts. This was investigated in two domains, comparing the predictive accuracy of the entropy-based measures to the Flesch-Kincaid readability formula (Kincaid et al., 1975).

5.1 Key Findings

In all four test corpora, the three readability measures showed significant increases in difficulty score at higher labelled difficulty levels. This is a strong indication that all three measures are, in fact, modelling text features which contribute to reading difficulty. Where measured, there was also significant positive correlation amongst all measures. The strongest such correlation was consistently found between the Flesch-Kincaid readability grade and the time-free model score.

In the domain of educational resources, the Flesch-Kincaid readability grade was the most accurate predictor of labelled text difficulty. Of the two entropy-based measures, the time-free model score was a better predictor of difficulty grade in this domain than the time-limited score, which appeared to be heavily dependent on text length. This is consistent with the strength of correlation between time-free score and Flesch-Kincaid grade: Flesch-Kincaid grade is the best predictor in this domain, so we would expect the entropy-based measure which correlates more strongly with Flesch-Kincaid grade to be a better predictor.

In the domain of manually simplified texts, the time-limited model score was the most accurate predictor of labelled text difficulty. In this domain, the time-free model and Flesch-Kincaid formula yielded similar predictive accuracy. This is contrary to findings for educational resources, where time-limited score was the least accurate predictor. Predictive accuracy was, for all three measures, better for manually simplified texts than educational resources. This suggests that the classification problem in this domain is 'easier', as text information content is held constant, while linguistic modifications have been used explicitly to alter the readability of the text. These modifications may be easier to detect using readability models than the linguistic and conceptual differences between educational resources for different school grades.

The nature of misclassifications in the textbook and OneStop corpora reveal that intermediate-grade texts are most likely to be misclassified in a three-class classification problem (Figures 3 and 7). Intermediate-grade texts are more likely to be misclassified with a lower, rather than higher, grade. This may be a result of the nonlinear relationship between true reading difficulty and labelled difficulty grade: the true difference in readability between difficulty grades may be greater between higher grades, which would be appropriate if older or more advanced students are able to progress faster.

Errors made by the entropy-based models was further investigated to identify potential causes of text misclassification. In the domain of educational resources, both models performed significantly worse than the Flesch-Kincaid formula, and were unable to correctly classify any of the intermediate (Grade 7) texts in the textbook corpus. There were also many misclassifications where Grade 6 texts were assigned Grade 8, and vice versa. The most extreme such cases were examined to inspect the Grade 6 texts with the highest reading difficulty scores, and the Grade 8 texts with the lowest predicted reading difficulty. As expected, the time-limited score's dependence on text length means that the Grade 6 text with the highest time-limited score is simply the longest Grade 6 text (3140 words), while the Grade 8 text with the lowest time-limited score is the shortest Grade 8 text (327 words). The Grade 6 text with the highest time-free score is a section describing civilizations in

Sumer, containing many place names and technical terms contributing to its high reading difficulty score (see Appendix F). A clear example of this is seen in the following sentence:

> By 3000 B.C.E. most Sumerians lived in powerful city-states like Ur, Lagash (LAY-gash), and Uruk (UH-ruhk).

The Grade 8 text with the lowest time-free score is a short introductory passage, consisting of short sentences to outline the chapter (see Appendix G). These short, simple sentences make the passage linguistically simple, despite its Grade 8 difficulty label:

> The great majority of African Americans lived in slavery. Harriet Powers was one of them. Powers was born into slavery in Georgia in 1837. Like many slaves, she grew up hearing Bible stories.

This explains the discrepancy between predicted reading difficulty and textbook grade for these texts, and highlights the limitations of using textbook grade as a label of reading difficulty without considering the actual contents of each textbook section.

In the domain of manually simplified texts, predictive accuracy was generally better than for educational resources, however there were still classification errors which can be explored. The 'Elementary level' OneStop articles with the highest time-free and time-limited scores were found to contain many quotations, names, and place names, contributing to their high reading difficulty scores. The 'Advanced level' OneStop article with the lowest time-free and time-limited score is an anecdotal article written in the first person with short, simple sentences. These factors explain the misclassification of these articles, as writing style and subject matter are detected by the readability models, but not reflected in the articles' labelled difficulty grade.

5.2 Limitations of Results

One key limitation of these results is the relatively small test corpora used, and the methods used to assign their labelled difficulty grades. In particular, the corpora of educational resources cover a variety of topics, with a labelled difficulty grade based on both the linguistic and conceptual difficulty of texts. Therefore, predicting difficulty grade requires some insight into the conceptual difficulty of the text, while the readability measures are only intended to predict linguistic difficulty. This is highlighted by the nature of misclassification errors made by the entropy-based models, as inspection of the misclassified articles reveals discrepancies between labelled difficulty grade and actual (subjective) reading difficulty.

It is also important to note that the entropy-based readability measures assessed in this study were trained by Xing et al. (2008) on textbook passages designed for Chinese L2 English students. The purpose of this investigation is to assess how well these models generalise to domains outside of their training-domain. However, it is impossible to extrapolate from these results the predictive accuracy of these measures if they were trained on data more similar to the test-domain. These results reveal only the potential of these measures for predicting reading difficulty, whereas training the model parameters on a wider domain or a domain closer to the intended test-domain may yield higher predictive accuracy.

This study provides a direct comparison of the entropy-based readability models against the Flesch-Kincaid readability formula. High correlation has been observed between different traditional readability formulae (Stajner et al., 2012), so this comparison suffices to assess the entropy-based against traditional formulae in general. More advanced readability models such as the Coh-metrix (Crossley et al., 2008) or CAREC model (Choi & Crossley, 2020; Crossley et al., 2019) are not publicly

available, so no direct comparison can be made to these state-of-the-art methods. Choi and Crossley (2020) evaluated a set of 'traditional' and 'new' readability models on the classification of Wikipedia and SimpleWiki articles, finding that CAREC outperformed the Flesch-Kincaid formula for predictive accuracy. Their methodology for training and testing their predictive models used a much larger corpus of articles than this study, and the discrepancy between their reported accuracy of the Flesch-Kincaid formula (82%) and the mean cross-validated accuracy in this study (94%) suggest that our methodologies are incompatible for comparison.

5.3 Open Questions

The entropy-based models assessed in this study are designed to match those developed by Xing et al. (2008), where model coefficients were trained on a corpus of Chinese L2 English textbooks. The results of this study indicate that these models generalise well to texts outside of their training domain; in particular, the time-limited model outperforms traditional formulae in domain of manually simplified texts, where information content is held constant. The predictive accuracy of both entropy-based models may be improved by re-training their coefficients on a broader training corpus, consisting of texts for a wider range of audiences, or on a specific domain to match the models' intended use, such as a corpus of manually simplified texts for a text-simplification application.

In this study, the relationship between model scores and labelled difficulty grade was assessed, however it is difficult to establish the relationship between labelled grade and true reading difficulty. In order to directly assess the relationship between model scores and true reading difficulty, data would need to be collected measuring participants' reading time and perceived difficulty for a text corpus. This data could be used to retrain the model coefficients to get predictors of reading time or perceived difficulty rather than labelled difficulty grade.

Another potential development for the entropy-based models would be to adapt the training corpus for different audiences, such as children, second-language learners, or domain experts. For example, a corpus of law textbooks could be added to the training corpus to compare the readability scores from these law-modified models to the scores from the original models. The predictive accuracy of the law-modified models could then be compared to the original models when predicting the perceived reading difficulty of texts for law students versus non-law students. If the law-modified models are a better predictor of the reading difficulty experienced by law students, this would indicate that the entropy-based readability models can effectively take into account the linguistic experiences of different audiences. This would be an important advantage for writers using readability measures to assess their material for specific audiences over other measures, which assume all readers will experience the same level of difficulty.

6. Conclusions

This investigation found significant correlation between entropy-based readability scores and graded reading difficulty in all test domains, indicating that these models are valid measures of reading difficulty in practically useful domains.

There was strong correlation between the time-free model score and the Flesch-Kincaid grade, which indicates that there is a correlation between word and sentence length (used in the Flesch-Kincaid formula) and average per-word and per-sentence entropy (used to compute the time-free model score). This is expected, as both are assumed to be measures of the amount of information conveyed by, and therefore difficulty in reading, the words and sentences in a text.

In the domain of educational resources, the best predictor of labelled difficulty was the Flesch-Kincaid formula. This is not unexpected, as traditional readability formulae have long been used by publishers to assess and amend their materials for different audiences (Dufty et al., 2006). In this context, it is difficult to determine the exact nature of this causal relationship, as easy-to-compute traditional readability formulae such as the Flesch-Kincaid formula are often used in the development process of educational resources. Therefore, the predictive accuracy of such readability formulae in this domain may be self-fulfilling to some degree. Of the entropy-based measures, the time-free model performed better than the time-limited model in this domain, consistent with the higher correlation between model score Flesch-Kincaid grade. The distribution of model scores revealed that time-limited scores mirror the distribution of text lengths, suggesting that, when information content is not held constant, the effect of text length on this measure may be overshadowing the effect of other linguistic factors.

In the domain of manually simplified texts, where information content is held constant, the time-limited score was the best predictor of difficulty level. These results indicate that the time-limited model is most useful for evaluating texts when information content is held constant, such as the output of text simplification tools.

The predictive accuracy of both entropy-based models could be improved by retraining their coefficients on data which more closely matches the intended test domain. It is also important to note that this investigation assessed the readability models by their ability to predict a labelled difficulty grade which is assumed to correlate with true difficulty. However, this is not always the case as perceived difficulty varies within labelled difficulty grades. Further investigation is required to assess the models' utility for predicting reading time or perceived difficulty, which may be better measures of 'true' reading difficulty.

References

Alexander, L. G., & Stoldt, P. H. (1972). *New Concept English* (Issue v. 2). Langenscheidt-Longman. https://books.google.co.uk/books?id=s6-UAAAACAAJ

Aluisio, S., Specia, L., Gasperin, C., & Scarton, C. (2010). Readability assessment for text simplification. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, 1–9.

Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully guage difficulty of reading materials? *Psychology in the Schools*, *51*(2), 198–215. https://doi.org/https://doi.org/10.1002/pits.21740

Besson, A. J., Kei, C., Jackson, B., Yeung, T. M., Deftereos, I., & Yeung, J. M. C. (2020). Patient information on the internet for surgical management of inflammatory bowel disease: is it good enough? *International Surgery Journal*, *8*(1), 12–18.

Bird, S., & Loper, E. (2020). *NLTK 3.5 documentation*. Natural Language Toolkit. https://www.nltk.org/

Bormuth, J. R. (1968). The Cloze readability procedure. *Elementary English*, *45*(4), 429–436. http://www.jstor.org/stable/41386340

Brownlee, J. (2019). *A gentle introduction to cross-entropy for machine learning*. Machine Learning Mastery. https://machinelearningmastery.com/cross-entropy-for-machine-learning/

Choi, J. S., & Crossley, S. A. (2020). *Assessing readability formulas: A comparison of readability formula performance on the classification of simplified texts*. EasyChair.

Crawford-Manning, F., Greenall, C., Hawarden, A., Bullock, L., Layland, S., Jinks, C., Protheroe, J., & Paskins, Z. (2020). Evaluation of quality and readability of online patient information on osteoporosis and osteoporosis drug treatment and recommendations for improvement. *Osteoporosis International*.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3–4), 541–561. https://doi.org/10.1111/1467-9817.12283

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, *27*(2), 37–54. http://www.jstor.org/stable/1473669

Dale, E., & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19–26. http://www.jstor.org/stable/41383594

DiMascio, C. (2019). *Determine the readability of a text with Python*. Level Up Coding. https://levelup.gitconnected.com/determine-the-reading-level-of-a-text-with-python-d2f9dccee6bf

Dubay, W. (2004). The principles of readability. *CA*, *92627949*, 631–3309.

Dufty, D. F., Graesser, A. C., Louwerse, M. M., & Mcnamara, D. S. (2006). Assigning Grade Levels to Textbooks: Is it just Readability? *The 28th Annual Conference of the Cognitive Science Society*, 1251–1256.

Ekwall, E. E., & Henry, I. B. (1968). How to find books children can read. *The Reading Teacher*, *22*(3), 230–232. http://www.jstor.org/stable/20196091

ESOL. (n.d.). *Reading for pleasure - Graded texts and short stories for intermediate English learners*. Retrieved November 26, 2020, from https://www.esolcourses.com/content/reading/intermediate-english-graded-readers.html

Fass, W., & Schumacher, G. M. (1978). Effects of motivation, subject activity, and readability on the retention of prose materials. *Journal of Educational Psychology*, *70*(5), 803–807. https://doi.org/10.1037/0022-0663.70.5.803

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A comparison of features for automatic readability assessment. *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, *2*, 276–284.

Flesch, R. (1948). *The Flesch reading ease formula*.

Francis, W. N., & Kucera, H. (1979). Brown Corpus Manual. *Letters to the Editor*, *5*(2), 7.

Government Digital Service. (2016, January). *Writing for GOV.UK - Content design: Planning, writing and managing content*. https://www.gov.uk/guidance/content-design/writing-for-gov-uk

Gray, W. S., & Leary, B. E. (1935). *What makes a book readable*. Univ. Chicago Press.

Gunning, R. (1952). The technique of clear writing. *New York, McGraw-Hill*.

Gunning, T. (2003). The role of readability in today's classrooms. *Topics in Language Disorders*, *23*, 175–189.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Hammond, M. (2006). Probabilistic Language Models. In *Mathematics of Language and Linguistics*. University of Arizona.

Hartley, J. (2016). Is time up for the Flesch measure of reading ease? *Scientometrics*, *107*(3), 1523–1526. https://doi.org/10.1007/s11192-016-1920-7

Hundt, M., Sand, A., & Siemund, R. (1998). *Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB')*. Albert-Ludwigs-Universität Freiburg.

Hundt, M., Sand, A., & Skandera, P. (1999). Manual of information to accompany the Freiburg-Brown Corpus of American English. *Freiburg: University of Freiburg*.

IEEE. (n.d.). *Write clearly and concisely*. IEEE Professional Communication Society. Retrieved January 6, 2021, from https://procomm.ieee.org/communication-resources-for-engineers/style/write-clearly-and-concisely/

Johansson, S., Atwell, E., Garside, R., & Leech, G. (1986). The Tagged LOB Corpus. *Norwegian Computing Centre for the Humanities*.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. *Probabilistic Linguistics*, *21*.

Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing*.

Kanter, H., Muscarelllo, T., & Ralston, C. (2008). Measuring the readability of software requirement specifications. *Information Systems Control Journal*, *1*.

Kincaid, J. P., Fishburne Jr, 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. Naval Technical Training Command Millington TN Research Branch.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, *10*(1), 62–102. https://doi.org/10.2307/747086

Kuang, Y. F., Lee, G., & Qin, B. (2020). Does government report readability matter? Evidence from market reactions to AAERs. *Journal of Accounting and Public Policy*, *39*(2), 106697. https://doi.org/10.1016/j.jaccpubpol.2019.106697

Lakretz, Y., Dehaene, S., & King, J. R. (2020). What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*, *22*(4), 446. https://doi.org/10.3390/E22040446

Lehavy, R., Li, F., & Merkley, K. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Accounting Review*, *86*(3), 1087–1115. https://doi.org/10.2308/accr.00000043

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In *Sentence Processing* (pp. 78–114). https://doi.org/10.4324/9780203488454

Ley, P., & Florio, T. (1996). The use of readability formulas in health care. *Psychology, Health and Medicine*, *1*(1), 7–28. https://doi.org/10.1080/13548509608400003

Linzen, T., & Jaeger, F. (2014). Investigating the role of entropy in sentence processing. *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics*, 10–18. https://doi.org/10.3115/v1/w14-2002

Marks, C. B., Doctorow, M. J., & Wittrock, M. C. (1974). Word frequency and reading comprehension. *Journal of Educational Research*, *67*(6), 259–262. https://doi.org/10.1080/00220671.1974.10884622

McLaughlin, G. H. (1969). SMOG grading - A new readability formula. *Journal of Reading*, *12*(8), 639–646. http://www.jstor.org/stable/40011226

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292–330. https://doi.org/10.1080/01638530902959943

OED Online. (2020). *readability, n.* Oxford University Press.

Plain English Campaign. (n.d.). *About us*. Retrieved October 30, 2020, from http://www.plainenglish.co.uk/about-us.html

Ratner, N. B., & Sih, C. C. (1987). Effects of gradual increases in sentence length and complexity on children's dysfluency. *Journal of Speech and Hearing Disorders*, *52*(3), 278–287. https://doi.org/10.1044/jshd.5203.278

Redish, J. (2000). Readability formulas have even more limitations than Klare discusses. *ACM Journal of Computer Documentation*, *24*(3), 132–137. https://doi.org/10.1145/344599.344637

Román, D. X., Briceño, A., Rohde, H., & Hironaka, S. (2016). Linguistic Cohesion in Middle-School Texts: A Comparison of Logical Connectives Usage in Science and Social Studies Textbooks. *Electronic Journal of Science Education*, *20*(6), 1–19.

Rubenstein, H., & Aborn, M. (1958). Learning, prediction, and readability. *Journal of Applied Psychology*, *42*(1), 28–32. https://doi.org/10.1037/h0039808

Schwartz, D., Sparkman, J. P., & Deese, J. (1970). The process of understanding and judgments of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, *9*(1), 87–93. https://doi.org/10.1016/S0022-5371(70)80013-5

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, *4*(1), 58–70.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, *165*(2), 259–298. https://doi.org/10.1075/itl.165.2.06sid

Sigurd, B., Eeg-Olofsson, M., & van Weijer, J. (2004). Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, *58*(1), 37–52. https://doi.org/10.1111/j.0039-3193.2004.00109.x

Sinyai, C., MacArthur, B., & Roccotagliata, T. (2018). Evaluating the readability and suitability of construction occupational safety and health materials designed for workers. *American Journal of Industrial Medicine*, *61*(10), 842–848. https://doi.org/10.1002/ajim.22901

Skwire, D., & Wiener, H. S. (1998). *Student's Book of College English: Rhetoric, Readings, Handbook*. Allyn and Bacon. https://books.google.co.uk/books?id=WEbglXP2n9IC

Sokolov, A. (2015). *Statistical Machine Translation*. https://www.cl.uni-heidelberg.de/courses/ss15/smt/scribe6.pdf

Stajner, S., Evans, R., Orasan, C., & Mitkov, R. (2012). What can readability measures really tell us about text complexity? *Proceedings of Workshop on Natural Language Processing for Improving Textual Accessibility*, 14–22.

Tater, K. C. (2021). Veterinary allergy information has lower health readability than human allergy information: a comparative analysis of allergy education materials for pets and people. *Veterinary Dermatology*.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

The Plain Writing Act, 1 (2010). http://www.gpo.gov/fdsys/pkg/PLAW-111publ274/pdf/PLAW-111publ274.pdf

Vajjala, S., & Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 297–304. https://doi.org/10.18653/v1/W18-0535

Vajjala, S., & Meurers, D. (2014). Readability assessment for text simplification. *ITL - International Journal of Applied Linguistics*, *165*(2), 194–222. https://doi.org/10.1075/itl.165.2.04vaj

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Semantic entropy in language comprehension. *Entropy*, *21*. https://doi.org/10.3390/e21121159

Wegner, M. V., & Girasek, D. C. (2003). How readable are child safety seat installation instructions? *Pediatrics*, *111*(3), 588–591. https://doi.org/10.1542/peds.111.3.588

Wikipedia. (n.d.-a). *Multiyear ranking of most viewed pages*. Wikipedia.Org. Retrieved February 25, 2021, from https://en.wikipedia.org/wiki/Wikipedia:Multiyear_ranking_of_most_viewed_pages

Wikipedia. (n.d.-b). *Simple Wikipedia*. Wikipedia.Org. Retrieved February 24, 2021, from https://simple.wikipedia.org/wiki/Main_Page

Worrall, A. P., Connolly, M. J., O'Neill, A., O'Doherty, M., Thornton, K. P., McNally, C., McConkey, S. J., & De Barra, E. (2020). *Readability of online COVID-19 health information: A comparison between four English speaking countries* (Version 3). https://doi.org/10.21203/rs.3.rs-30124/v3

Xing, F., Cheng, D., & Pu, J. (2008). A new approach to readability study based on information computing. *2008 International Conference on Advanced Language Processing and Web Information Technology*, 156–161. https://doi.org/10.1109/ALPIT.2008.37

Yan, S., & Jaeger, T. F. (2020). Expectation adaptation during natural reading. *Language, Cognition and Neuroscience*, *35*(10), 1394–1422. https://doi.org/10.1080/23273798.2020.1784447

Yeomans, L. (2009). *Evaluation of product documentation provided by suppliers of hand held power tools*.

Appendix A

List of Genres included in Training Corpus

I. Informative Prose
   - A. Press: Reportage
   - B. Press: Editorial
   - C. Press: Reviews
   - D. Religion
   - E. Skills and Hobbies
   - F. Popular Lore
   - G. Belles Lettres, Biography, Memoirs
   - H. Miscellaneous
   - J. Learned (Education)

II. Imaginative Prose
   - K. General Fiction
   - L. Mystery and Detective Fiction
   - M. Science Fiction
   - N. Adventure and Western Fiction
   - P. Romance and Love Story
   - R. Humour

(Francis & Kucera, 1979)

Appendix B

List of Textbooks Analysed in Textbook Corpus

CPO Science. (2007). *Focus on Earth Science* (CA Ed.). Cambridge, MA: Cambridge Physics Outlet.

CPO Science. (2007). *Focus on Life Science* (CA Ed.). Cambridge, MA: Cambridge Physics Outlet.

CPO Science. (2007). *Focus on Physical Science* (CA Ed.). Cambridge, MA: Cambridge Physics Outlet.

Glencoe/McGraw Hill. (2006). *Discovering Our Past: Ancient Civilizations* (California ed.). Columbus, OH: Glencoe/McGraw-Hill.

Glencoe/McGraw Hill. (2006). *Discovering Our Past: Medieval and Early Modern Times* (California edition). Columbus, OH: Glencoe/McGraw-Hill.

Glencoe/McGraw Hill. (2006). *Discovering Our Past: The American Journey to World War I.* (California ed.). Columbus, OH: Glencoe/McGraw-Hill.

Glencoe/McGraw-Hill. (2007). *Focus on Earth Science (California edition).* Columbus, OH: Glencoe/McGraw-Hill.

Glencoe/McGraw-Hill. (2007). *Focus on Life Science (California edition).* Columbus, OH: Glencoe/McGraw-Hill.

Glencoe/McGraw-Hill. (2007). *Focus on Physical Science (California edition).* Columbus, OH: Glencoe/McGraw-Hill.

Holt, Rinehart, and Winston Inc. (2007). *Earth Science* (CA Ed.). Boston, MA: Houghton Mifflin Harcourt Publishing Company.

Holt, Rinehart, and Winston Inc. (2007). *Life Science* (CA Ed.). Boston, MA: Houghton Mifflin Harcourt Publishing Company.

Prentice Hall. (2006). *Ancient Civilizations* (CA Ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Prentice Hall. (2008). *Focus on Earth Science* (CA Ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Prentice Hall. (2008). *Focus on Life Science* (CA Ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Prentice Hall. (2008). *Focus on Physical Science* (CA Ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Prentice Hall. (2006). *Medieval and Early Modern Times* (CA Ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

TCI. (2011). *History Alive! The Ancient World*. Palo Alto, CA: Teachers' Curriculum Institute.

TCI. (2011). *History Alive! The Medieval World and Beyond.* Palo Alto, CA: Teachers' Curriculum Institute.

TCI. (2011). *History Alive! The United States Through Industrialism*. Palo Alto, CA: Teachers' Curriculum Institute.

(Román et al., 2016)

Appendix C

List of Articles Analysed in Wikipedia/SimpleWiki Corpus

United States
Donald Trump
Barack Obama
India
World War II
Michael Jackson
Elizabeth II
United Kingdom
Lady Gaga
Eminem
Sex
Adolf Hitler
Game of Thrones
Cristiano Ronaldo
World War I
The Beatles
Justin Bieber
Canada
Steve Jobs
Freddie Mercury
Kim Kardashian
The Big Bang Theory
Australia
Michael Jordan
Lionel Messi
Stephen Hawking
Dwayne Johnson
Darth Vader
Star Wars
Miley Cyrus
China
Taylor Swift
Academy Awards
Lil Wayne
Abraham Lincoln
Japan
Germany
How I Met Your Mother
Selena Gomez
Harry Potter
September 11 attacks
Johnny Depp
New York City
Rihanna
Kobe Bryant
Elon Musk
Russia
Albert Einstein
The Walking Dead (TV series)
LeBron James
Kanye West
Tupac Shakur
Leonardo DiCaprio
Angelina Jolie
France
Breaking Bad
Chernobyl disaster
Earth
Mila Kunis
Vietnam War
John F. Kennedy
Mark Zuckerberg
Arnold Schwarzenegger
Tom Cruise
Pablo Escobar
Scarlett Johansson
William Shakespeare
Bible
Jennifer Aniston
COVID-19 pandemic
Bill Gates
Will Smith
Ariana Grande
Nicki Minaj
Glee (TV series)
Muhammad Ali
Katy Perry
Ted Bundy
Charles Manson
John Cena
Keanu Reeves
Queen Victoria
Singapore
Israel
Illuminati
Sexual intercourse
Bruce Lee
Elvis Presley
Marilyn Monroe
London
Adele
Doctor Who
Prince (musician)
Brad Pitt
Periodic Table
Joe Biden
Heath Ledger
American Civil War
Jay-Z
David Bowie

(Wikipedia, n.d.-a)

Appendix D

Summary Statistics and Statistical Test Results for Textbook Corpus Scores

Table 1

Summary Statistics in each Textbook Grade

Grade	Count	Length	Flesch-Kincaid	Time-Free Model Score	Time-Limited Model Score
6	634	μ = 904 σ = 623	μ = 7.68 σ = 0.94	μ = 239.90 σ = 25.14	μ = 486.64 σ = 234.92
7	576	μ = 1032 σ = 610	μ = 8.15 σ = 0.99	μ = 240.47 σ = 23.03	μ = 534.32 σ = 230.78
8	600	μ = 1205 σ = 711	μ = 8.62 σ = 1.17	μ = 251.34 σ = 26.10	μ = 596.70 σ = 258.54

Table 2

Levene's test for Equality of Variances Across Textbook Grades

Measure	F Value
Flesch-Kincaid	20.32*
Time-Free	2.84
Time-Limited	2.88

*p<.05

Table 3

Welch's Test for Difference in Group Means

Measure	F Value
Flesch-Kincaid	122.39**
Time-Free	38.57**
Time-Limited	30.50**

**p<.01

Appendix E

Summary Statistics and Statistical Test Results for OneStop Corpus Scores

Table 1

Summary Statistics in each OneStop Difficulty Grade

Grade	Count	Length	Flesch-Kincaid	Time-Free Model Score	Time-Limited Model Score
1 – Elementary	189	μ = 604.63 σ = 117.43	μ = 9.14 σ = 1.59	μ = 321.77 σ = 49.06	μ = 430.67 σ = 54.98
2 – Intermediate	189	μ = 758.54 σ = 132.26	μ = 10.07 σ = 1.65	μ = 332.24 σ = 47.22	μ = 490.66 σ = 56.48
3 – Advanced	189	μ = 926.44 σ = 183.08	μ = 12.75 σ = 2.27	μ = 433.96 σ = 74.61	μ = 638.06 σ = 83.27

Table 2

Levene's test for Equality of Variances Across OneStop Difficulty Grades

Measure	F Value
Flesch-Kincaid	12.67*
Time-Free	23.36*
Time-Limited	20/76*

*p<.05

Table 3

Friedman Test for Difference in Distributions Across OneStop Difficulty Grades

Measure	Chi-Squared
Flesch-Kincaid	322.51**
Time-Free	287.66**
Time-Limited	351.66**

**p<.01

Appendix F

Grade 6 Textbook Passage with the Highest Time-Free Score

TCI. (2011). *History Alive! The Ancient World*. Palo Alto, CA: Teachers' Curriculum Institute. Unit 1. Chapter 5. Section 2.

Section 2: Characteristics of Civilization

Sumer was a challenging place to live.

It had hot summers, little rain, and rivers that flooded the plains in the spring.

Yet the Sumerians were able to overcome these challenges.

They built complex irrigation systems and large cities.

By 3000 B.C.E. most Sumerians lived in powerful city-states like Ur, Lagash (LAY-gash), and Uruk (UH-ruhk).

But what did the Sumerians do to create a civilization?

To answer this question, we need to examine what civilization means.

What characteristics make a society into a civilization?

Historians name several such characteristics, including these:

- A stable food supply, to ensure that the people of a society have the food they need to survive.
- A social structure with different social levels and jobs.
- A system of government, to ensure that life in the society is orderly.
- A religious system, which involves both a set of beliefs and forms of worship.
- A highly developed way of life that includes the arts, such as painting, architecture, music, and literature.
- Advances in technology.
- A highly developed written language.

Did Sumer have these characteristics?

Let's find out what the evidence can tell us.

Appendix G

Grade 8 Textbook Passage with the Lowest Time-Free Score

TCI. (2011). *History Alive! The United States Through Industrialism*. Palo Alto, CA: Teachers' Curriculum Institute. Unit 6. Chapter 20. Section 1.

Section 1: Introduction

By 1850, the population of the United States had grown to just over 23 million.

This figure included 3.6 million African Americans.

The great majority of African Americans lived in slavery.

Harriet Powers was one of them.

Powers was born into slavery in Georgia in 1837.

Like many slaves, she grew up hearing Bible stories.

In her quilts, she used animals and figures from Africa and the United States to illustrate those stories, along with scenes from her life.

Hidden in her images were messages of hope and freedom for slaves.

Not all African Americans were slaves.

By mid-century, there were about half a million free blacks as well.

Many were former slaves who had escaped to freedom.

Whether African Americans lived in slavery or freedom, discrimination (unequal treatment) shaped their lives.

Throughout the country, whites looked down on blacks.

Whites ignored the contributions blacks made to American life.

They thought of the United States as their country.

Such racist thinking later prompted African American scholar and reformer W. E. B. Du Bois to ask, Your country?

How came it to be yours?

Before the Pilgrims landed, we were here.

Here we brought you our three gifts and mingled them with yours; a gift of story and song, soft, stirring melody in an unmelodious land; the gift of sweat and brawn to beat back the wilderness and lay the foundations of this vast economic empire the third, a gift of the Spirit.

In this chapter, you will explore how African Americans faced and endured discrimination and slavery in the mid-1800s.

You will also learn more about the gifts that African Americans brought to America.

B122164_Dissertation_10661-Words

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44

PAGE 45

PAGE 46