A Readability Level Prediction Tool for K-12 Books

Joel Denning
Computer Science Department, Brigham Young University, Provo, Utah 84602, U.S.A.
Email: joeldenning@gmail.com

Maria Soledad Pera
Computer Science Department, Brigham Young University, Provo, Utah 84602, U.S.A.
Email: mpera@cs.byu.edu

Yiu-Kai Ng
Computer Science Department, Brigham Young University, Provo, Utah 84602, U.S.A.
Email: ng@compsci.byu.edu

Abstract
The readability level of a book is a useful measure for children and teenagers (teachers, parents, and librarians, respectively) to identify reading materials suitable for themselves (their K-12 readers, respectively). Unfortunately, majority of published books are assigned a readability level range, such as K-3, instead of a single readability level for their intended readers, by professionals, which is not useful to the end-users who look for books at a particular grade level. This leads to the development of readability formulas/analysis tools. These formulas/tools, however, require at least an excerpt of a book to estimate its readability level, which is a severe constraint due to copyright laws that often prohibit book content from being made publicly accessible. To alleviate the text constraint imposed on readability analysis on books, we have developed TRoLL, which relies heavily on metadata of books that is publicly and readily accessible from reputable book-affiliated online sources, besides using snippets of books, if they are available, to predict the readability level of books. Based on a multi-dimensional regression analysis, TRoLL determines the grade level of any book instantly, even without a sample text in the book, which is its uniqueness. TRoLL handles the entire spectrum of readability levels, i.e., the well-known K-12 grade-level range. Unlike its counterparts, TRoLL explicitly considers the topical suitability of (the content of) a book in determining its readability level, which is one of the unique features of TRoLL as a readability-level prediction tool. Furthermore, TRoLL is a significant contribution to the educational community, since its computed book readability levels can (i) enrich K-12 readers’ book selections and thus can enhance their reading for learning experience, and (ii) aid parents, teachers, and librarians in locating reading materials suitable for their K-12 readers, which can be a time-consuming and frustrating task that does not always yield a quality outcome. Empirical studies conducted using a large set of K-12 books have verified the prediction accuracy of TRoLL and demonstrated its superiority over existing well-known readability formulas/analysis tools.

Introduction
As reading is an essential skill, which can have significant impact on a youth’s educational and future career development (Robinson et al., 2011), it is imperative to encourage children to read and learn starting from an early age. Reading for learning, however, cannot take place unless readers can accurately and
efficiently decode, i.e., comprehend, the words in a text (Oakhill and Cain, 2012). During the last century, educators and researchers have dedicated resources to develop readability assessment tools/formulas which quantify the degree of difficulty in understanding a text (Benjamin, 2012; Feng et al., 2010).

Traditional readability formulas, such as Flesch-Kincaid (Reading Ease) (Kincaid et al., 1975), simply perform a one-dimensional analysis on a text based on shallow features, such as the average number of syllables per word (words per sentence, respectively), the average sentence length, and vocabulary lists, which might not precisely capture the complexity of a text. More recently-developed readability formulas have gone beyond shallow features and rely on natural language processing tools to examine complex linguistic features on a text (Feng et al., 2010). All of these formulas, however, require a given (snippet of a) text in order to determine its readability level (i.e., grade level), which is a constraint if applied to books, since even an excerpt of a book is not always freely accessible due to copyright laws. The same constraint affects Lexile Framework (Smith et al., 1989) and Advantage-TASA Open Standard for Readability (ATOS) (School Renaissance Institute Inc., 2000), two widely-used readability analysis tools these days specifically developed for analyzing the readability level of books.

To address the deficiencies of the designs of existing readability formulas/analysis algorithms, we propose a tool for regression analysis of literacy levels, denoted TRoLL, which considers metadata of books publicly accessible from reputable online sources, in addition to snapshots of books only if they are available, to predict the grade level of any book. To determine the grade level of a book \( Bk \), TRoLL extracts from well-known sources, such as WorldCat.org, (i) an excerpt from \( Bk \) (if it is available online) to analyze various shallow features, determine its subject area established by the US Curriculum, and examine different grammatical concepts in \( Bk \); (ii) the subject headings assigned to \( Bk \); and (iii) the targeted audience level of \( Bk \) and its (first) author, besides the subject headings and audience levels of books written by the author.

TRoLL predicts the grade level of K-12 books. The grade levels can serve as a guideline for young readers to select books by themselves, which is a valuable, and often overlooked, tool, since “when students choose books that match their interests and level of reading achievement, they gain a sense of independence and commitment and they are more likely to complete, understand, and enjoy the book they are reading” (Milone, 2012). The grade levels predicted by TRoLL on (non-)fictional books and textbooks can also be used as a guidance for parents, teachers, and librarians in locating reading materials suitable for their K-12 readers.

TRoLL is unique, since it can predict the grade level of a book instantly, even if its sample text is unavailable online. TRoLL performs a multi-dimensional analysis on the metadata/content of books and their authors to accurately predict the readability level of books. Unlike other readability formulas/tools, such as Lexile, which predict the difficulty of a text based on their own readability-level scales, TRoLL predicts the grade level of a book, a measure preferred by teachers/librarians, given that grade levels are easy to understand and use when communicating with students/patrons (Renaissance Learning, 2011).

The main contribution of TRoLL is in its development as a tool that can determine the grade level of books on-the-fly, requiring solely on publicly available information on books and without involving human experts. This task cannot be accomplished by existing text-based readability formulas nor the popular Lexile or ATOS that offer readability measures for only a small fraction of published books and require direct involvement from their developers in order to generate the readability level of books that have yet

---

2 Davison and Kantor (Davison and Kantor, 1982) claim that “nonsense text” can be classified as easy-to-read by traditional readability formulas if it contains frequently-used, short words organized into brief sentences.

3 We have empirically verified that by considering only the first author of a K-12 book, the processing time of TRoLL is minimized without affecting its accuracy in predicting the grade level of the book. This is expected, since among the hundreds of thousands of K-12 books we have sampled at ARbookfind.com and Scholastic.com, less than 10% are written by multiple authors.
to be analyzed (Benjamin, 2012). As a by-product of our work, we have created a dataset consisting of more than 18,000 books with their respective grade level ranges defined by their corresponding publishers. Given the difficulty in obtaining large-scale datasets on books for training/testing a grade-level prediction tool on books (Tanaka-Ishii et al., 2010), the constructed dataset is an asset to the research community.

The remaining of this paper is organized as follows. In the “Related Work” section, we compare the design methodologies of existing readability-level prediction formulas/tools with TRoLL. In the “Our Readability Analysis Tool” section, we present the detailed design of TRoLL. In the “Experimental Results” section, we analyze the results of the empirical studies conducted to validate the correctness of TRoLL and include the comparisons of its performance with a number of well-known readability formulas/tools. In the “Conclusions and Future Work” section, we give a concluding remark and provide directions for future research work.

Related Work

For almost a century, readability formulas/analysis tools have been developed to determine the readability level or degree of difficulty of a text, resulting in hundreds of them (DuBay, 2004). Traditional formulas, including Flesch-Kincaid (Kincaid et al., 1975) and Gunning Fog (Index) (Gunning, 1952), are based on shallow features. The Flesch Reading Ease Readability Formula (Flesch, 1948) employs a 100-point scale to predict the “difficulty” of a text such that the lower its score, the harder it is to understand. Flesch-Kincaid (Kincaid et al., 1975), which is an enhancement of the Flesch Reading Ease formula, predicts the grade level of a text. Even though Flesch-Kincaid is similar to its predecessor in terms of considering word/sentence lengths for readability prediction, it applies a different weighting scheme in prediction. Besides considering the number of sentences in a text, Simple Measure of Gobbledygook, a readability formula better known as SMOG formula (McLaughlin, 1969), also examines the number of polysyllables words (i.e., words with more than two syllables) per every thirty sentences in the text to determine its grade level, which ranges from 5 to 18. Gunning Fog, on the other hand, computes an index score based on the average length of sentences in a text and the number of “complex” words (i.e., words with three or more syllables) within every 100 words. As stated in (Kodom, 2013), this measure is based on the premise that texts with short sentences consisting of simple words are easy to understand, which explains why “the ideal readability score for the Fog index is 7 or 8. Anything above 12 is too hard for most people to read” (Kodom, 2013).

Other popular formulas, which predict the readability level of a text by analyzing the average number of sentences/(difficult) words/syllables per word in the text, include Dale-Chall (Chall, 1995), Forcast (Caylor et al., 1973), Fry (Fry, 1968), PSK (Power et al., 1958), and Spache (Spache, 1953). These formulas, however, only provide a rough estimation of the difficulty of a text and thus are not always reliable (Benjamin, 2012; Feng et al., 2010). Lexile (Smith et al., 1989) and ATOS (School Renaissance Institute Inc., 2000), two well-known readability analysis tools, are based upon traditional readability features. While the former compares words in a text with 600 million words in the Lexile corpus to establish the semantic difficulty (i.e., word frequency) and syntactic complexity (i.e., sentence length) of the text, the latter considers word length, sentence length, and grade level of words, in addition to book length, i.e., word count, when it is applied to books. As indicated in a recent study on readability formulas (Begeny and Greene, 2014), a number of the aforementioned formulas/tools are applicable to predict a certain range of grade levels, e.g., “below fourth grade”, or generate a score (rather than a grade level), which makes it difficult to interpret its intended representation due to the lack of correlation between the

---

4The difficulty of a text is determined by the average sentence length and average number of syllables per word.

5Difficult terms are identified using a pre-defined list of words considered by the corresponding readability formula.
predicted value and the corresponding reading ability of an individual. TRoLL, however, computes an easy-to-interpret value that corresponds to a K-12 grade level.

Besides the formulas/tools listed above, new readability analysis approaches based on linguistic features have been developed (Collins-Thompson and Callan, 2004; Graesser et al., 2004; Heilman et al., 2008; Schwarm and Ostendorf, 2005). Coh-Metrix (Graesser et al., 2004) uses lexicons, part-of-speech classifiers, latent semantic analysis, and syntactic parsers, to name a few, to determine the difficulty of a text, which is influenced by cohesion relations and language/discourse characteristics. Collins-Thompson and Callan (Collins-Thompson and Callan, 2004) combine multiple statistical language models, which capture patterns of word usage in different grade levels, using a Na"ive Bayes classifier to estimate the most probable grade level of a text. Schwarm and Ostendorf (Schwarm and Ostendorf, 2005) apply support vector machines on various features extracted from statistical language models, along with shallow features and features derived from analyzing the syntactic structure of texts, to determine the readability level of a text $T$. Heilman et al. (Heilman et al., 2008) consider lexical and grammatical features derived from syntactic structures to analyze the difficulty of $T$. Regardless whether existing formulas are based on shallow and/or lexical features, these formulas do not consider the fact that (i) shortest words are not always simpler, (ii) least “difficult” words are not often easier to understand, and (iii) shortest sentences are not necessarily clear or most readable (Klare and Buck, 2013). In addition, these formulas do not always conduct an in-depth analysis on a book’s (i) content, which determines the suitability of the topics addressed in a text, and (ii) context, which estimates the satire in a text. More importantly, a number of existing readability-prediction formulas/tools are not “sensitive enough in predicting the earliest stages of reading” (Klare and Buck, 2013; Mesmer, 2007). TRoLL, on the other hand, combines text-based, as well as topical-based, features to determine an insightful/sophisticated/comprehensive readability level of a text. (For a detailed discussion on commonly-used features for assessing the readability of a text, see (Feng et al., 2010).)

Qumsiyeh and Ng (Qumsiyeh and Ng, 2011) and Ma et al. (Ma et al., 2012) have recently developed their own readability assessment tools. ReadAid (Qumsiyeh and Ng, 2011) performs an in-depth analysis beyond exploring the lexicographical and syntactical structures of an excerpt of a book by considering the authors of the book along with topic(s) covered in the book. Besides examining text-based features, SVM-Ranker (Ma et al., 2012) considers visually-oriented features (such as the average font size and ratio of annotated image rectangle area to page area) and adopts a rank-based strategy, as opposed to the commonly-employed classification/regression approaches, to determine the grade level of a book.

ReadAid (Qumsiyeh and Ng, 2011), ATOS (School Renaissance Institute Inc., 2000), and SVM-Ranker (Ma et al., 2012), along with the aforementioned readability formulas, either partially or fully depend on the availability of at least a sample of a text to compute its grade level, which is a severe constraint, since text in a book is not always freely accessible, either online or in a hard copy, due to copyright laws. TRoLL bypasses this constraint by using publicly available metadata on books to accomplish its task. (See (Benjamin, 2012; Begeny and Greene, 2014; DuBay, 2004) for an in-depth discussion of other existing readability formulas.)

**Our Readability Analysis Tool**

To alleviate the reliance of existing readability formulas/analysis tools on the text of a book, and to improve upon one-dimensional approaches towards determining readability levels of books, we introduce TRoLL, a sophisticated readability analysis tool that can operate without book content, i.e., sample text. Given a unique identifier of a book $Bk$, which is either its ISBN or its title and (first) author, TRoLL either retrieves the pre-computed readability level of $Bk$, if it has already been determined by TRoLL,
or calculates its readability level on the fly using a multiple linear regression model which analyzes publicly accessible information on $Bk$ that are offered by professional providers, who are either government or educational agents, and can be extracted online. Examples of such providers include the Library of Congress, the Online Computer Library Center (OCLC), and Open Library. These freely accessible information sources often include metadata, such as subject headings assigned to $Bk$, and occasionally include the target audience and/or the partial/full text of $Bk$. The overall readability prediction process of TRoLL is depicted in Figure 1.

Multiple Regression Analysis

To predict the readability level of a book $Bk$, TRoLL employs multiple linear regression analysis (Wooldridge, 2009), which is a classical statistical technique for building estimation models (Tan et al., 2009). The analysis accounts for the influence of multiple contributing factors, which are derived from metadata and/or content of $Bk$, to estimate the readability level of $Bk$ using the following equation:

$$ y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n $$

where $y$ is the dependent variable, which is the predicted readability level of $Bk$, $\beta_0$ is the intercept parameter, $\beta_1, \ldots, \beta_n$ are the coefficients of regression, $X_i$ ($1 \leq i \leq n$) is an independent variable (predictor), and $n$ is the number of predictors in the regression analysis (Wooldridge, 2009).

In Equation 1, each unknown parameter, i.e., the intercept and coefficients of regression, which is required to predict the readability level of a book by TRoLL, is estimated through a one-time training process using the Ordinary Least Squares method (Wooldridge, 2009) and the BookRL-RA training dataset (to be introduced in the Experimental Results section). Each book $b$ in BookRL-RA is represented as a vector of the form $<b_1, \ldots, b_{54}, r>$, where $b_i$ is the (value of the) the $i^{th}$ predictor ($1 \leq i \leq 54$) computed for $b$, and $r$ is the target, i.e., the known readability level for $b$ in our case. (The fifty-four predictors included in the regression model of TRoLL are defined in subsequent sections, whereas the

---

6http://www.loc.gov
7http://www.worldcat.org
8http://www.openlibrary.org
9We are aware that book repositories, such as WorldCat and OpenLibrary, occasionally archive more than one record on a given book $B$. Hence, whenever we refer to “metadata” of $B$, we mean the combination of the metadata extracted from all the archived records of $B$, providing that they are book records, i.e., records related to audiobooks, videos, or CDs are ignored. Moreover, records of $B$ are determined by the ISBNs of $B$, which can be obtained using publicly accessible APIs, such as “ThingISBN” offered by LibraryThing (http://www.librarything.com/api).
target readability level of a book is determined by its publisher and included in BookRL-RA.) Since publishers usually suggest a range of readability levels for each of their published books, such as grades 3-6, TRoLL considers the average grade level of the range as the target grade level of a book to avoid any bias by assigning books to their lowest or highest grade levels in the ranges during the regression training.

The Ordinary Least Squares method calculates the residual of each book \( b \) in BookRL-RA, which is the difference between the target readability level of \( b \) and the readability level of \( b \) predicted using the (values of the) predictors in the vector representation of \( b \) and Equation 1. Unknown parameters are estimated by minimizing the sum of squared distances between residuals of books in BookRL-RA.

### Analyzing Book Content

According to (Chen, 2012), only 7.7% of books in the OCLC database, which is a worldwide library cooperative that offers services to improve access to the world’s information, are linked to their partial or full content. We found similar results among the 7,142 books in the BookRL-RA dataset: only 5% of them include their partial or full content. Despite the low percentage of books with available content online, TRoLL utilizes the content of a book, if it is available, in predicting the readability level of the book.

Available online content of a book is either a snippet of less than five pages of the book, a preview of one or more of its chapters, or its full text (Chen, 2012). The analysis of book content is the basis for a number of TRoLL predictors, which rely on (i) textual features considered by traditional readability formulas, (ii) the grammar of its content, or (iii) subject areas addressed in the book. When calculating the values of these predictors, we only consider from the first up till the last sentence that includes the first 2,500th characters\(^{10}\) of the content of a book in order to improve the efficiency of TRoLL. We detail the analysis of the content of a book below.

### Predictors Based on Features Used by Traditional Readability Formulas

Existing widely-accepted readability formulas, such as Flesch-Kincaid (Kincaid et al., 1975), Coleman-Liau (Index) (Coleman, 1975), Spache (Readability Index) (Spache, 1953), Gunning Fog (Gunning, 1952), and SMOG (Index) (McLaughlin, 1969), seek to combine, through a mathematical formula, several textual features to compute the readability level of a text. We do not use any of these readability formulas as a TRoLL predictor, since there is no consent on which readability formula is the most accurate. Instead, we consider the features based on vocabulary and the count of syllables that are commonly used by traditional readability formulas as predictors so that TRoLL is not biased towards any particular readability formula.

TRoLL considers seven traditional textual features used in readability formulas: the count of (i) long words (with more than six letters), (ii) sentences, (iii) total words, (iv) letters, (v) syllables, (vi) words with three or more syllables, and (vii) unique unfamiliar words (Spache, 1953). Since the length of the text, i.e., the total number of characters, available online is different for each book, we normalize these counts to the length of the text.

**Example 1** Consider the book “A Wrinkle in Time,” denoted \( Bk_1 \), written by Madeleine L’Engle, which tells the story of a 14-year-old, Meg Murry, who lives a normal life until she enters a science fiction/fantasy world in which she goes on adventures. Its publisher suggests the target readers for the book to be in grades 5-7. Based on the first twenty pages of \( Bk_1 \)’s text that are publicly available (a sample of which is shown

\(^{10}\)The number of characters examined by TRoLL corresponds to the average number of words, i.e., 300 words, often examined by well-known readability formulas, which include Flesch-Kincaid, Fry, and Lexile, to determine the readability level of a text (Friedman and Hoffman-Goetz, 2006).
Figure 2: A sample of the text in “A Wrinkle in Time” in which long words are in **bold**, unfamiliar words (determined by traditional readability formulas) are *italicized*, and words with three or more syllables are *underlined*

in Figure 2), TRoLL analyzes a snippet with the first 2,639 characters (including the last sentence with the 2,500th character) and calculates (the values of) the following predictors: Count of long words = \( \frac{81}{2,639} = 0.031 \), Count of sentences = \( \frac{30}{2,639} = 0.015 \), Count of total words = \( \frac{475}{2,639} = 0.032 \), Count of letters = \( \frac{2,018}{2,639} = 0.765 \), Count of syllables = \( \frac{635}{2,639} = 0.241 \), Count of words with three or more syllables = \( \frac{29}{2,639} = 0.011 \), Count of unique unfamiliar words = \( \frac{86}{2,639} = 0.033 \).

**Grammar Predictors on Book Content**

TRoLL examines grammatical constructions, as defined by the US curriculum and shown in Table 1, to compute the values of grammar predictors. These predictors reflect the complexity of the (i) writing style, (ii) organization of the sentences, and (iii) grammatical constructs found in a text. The analysis of the grammar of textual content in a book \( Bk \) is somewhat more profound, due to advances in natural language processing, such as the Stanford NLP Parser (De Marneffe, 2006), than the analysis used in Flesch-Kincaid (Kincaid et al., 1975), Coleman-Liau (Coleman, 1975), Spache (Spache, 1953), Gunning Fog (Gunning, 1952), SMOG (McLaughlin, 1969), and other readability formulas.

There are two types of predictors created using grammatical constructions: simple and parse-tree. For *simple* grammatical concepts (listed in Table 1), which are easily measured, TRoLL simply counts their occurrences per sentence in the text of a book \( Bk \). When a grammatical concept is *more difficult* to find and count, TRoLL employs the Stanford Parser (De Marneffe, 2006) to parse the text into *parse trees*. Hereafter, TRoLL counts the occurrences of a grammatical structure per *parse tree* and normalizes the frequency of occurrence of the grammatical structures so that they are comparable regardless of the length of the text.

The grammatical predictors offer an in-depth analysis on the grammar of the textual content of \( Bk \), which are valuable to the regression analysis conducted by TRoLL.

**The Subject Area Predictor on Book Content**

TRoLL takes advantage of the mapping established by the US curriculum between *subject areas* and *grade levels* and exposes the subject area covered in a book to predict its readability level. A subject area is a specific topic specified in the US curriculum that is taught to students at a particular grade in the US public school system. For example, multiplication is taught at the 3rd grade, whereas geometry at the 10th. TRoLL pre-defines fifty-five distinct subject areas to be considered. These subject areas (and their corresponding grade levels) were inferred from the K-12 curriculum posted under Elkhart Community
Table 1: List of predictors used by TRoLL

<table>
<thead>
<tr>
<th>Predictors Based on Content (37)</th>
<th>Predictors Based on Traditional Text Features (?)</th>
<th>Predictors Based on Grammatical Constructions (29)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count of long words</td>
<td>Count of sentences</td>
<td>Count of words with three or more syllables</td>
</tr>
<tr>
<td>Count of syllables</td>
<td>Count of total words</td>
<td>Count of unique unfamiliar words</td>
</tr>
<tr>
<td>Count of sentences</td>
<td>Count of letters</td>
<td>Count of words with three or more syllables</td>
</tr>
<tr>
<td>Count of total words</td>
<td></td>
<td>Count of unique unfamiliar words</td>
</tr>
<tr>
<td>Count of letters</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Simple
- Common prefixes (un-, re-, pre-, in-, de-)
- Conjunctions (and, but, or)
- Determiners
- Irregular vowel combinations, spelling, phonetics (boot, soil, trout)

### Parse-tree Based
- Personal pronouns (him, her, it)
- Possessive nouns
- Prepositions
- Syncategorematic words (like, as, to, if, all)

### Adverbial phrases
- Adverbials
- Possessive pronouns
- Comparative verbs
- Participles
- Verbs of deduction
- Participles
- Participles
- Participles
- Participles

### Interrogative sentences
- Adverbial phrases
- Possessive pronouns
- Comparative verbs
- Participles
- Verbs of deduction
- Participles
- Participles
- Participles
- Participles

### Adverb phrases
- Adverbial phrases
- Possessive pronouns
- Comparative verbs
- Participles
- Verbs of deduction
- Participles
- Participles
- Participles
- Participles

### Compare
- Subject headings
- Frequency distribution predictors: mean, median, lowerBound, upperBound (4)
- Frequency distribution predictors within one standard deviation: SD_mean, SD_median, SD_lowerBound, SD_upperBound (4)
- Number of previously encountered subject headings
- Ratio of previously encountered subject headings assigned to books written by an author
- Median of readability levels paired with subject headings assigned to books written by an author

### Book audience level
- Average author’s audience level
- Minimum author’s audience level
- Maximum author’s audience level

To determine the subject area of a book $B_k$, TRoLL first analyzes (an excerpt of) its content by using a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003), which is a generative probabilistic model that represents documents as random mixtures over (latent) topics such that each topic is characterized by a distribution over words. To train a LDA model, we pre-defined the number of latent topics to be fifty-five, which match the number of subject areas considered by TRoLL, and applied JGibbLDA, a Java implementation of LDA, on 5,500 training documents randomly chosen from Wikipedia.org. Note that stopwords in the documents were removed and the remaining words were reduced to their grammatical root using the well-known Porter stemmer. During the training process, the LDA model estimates the probability distribution of words in latent topics (topics in documents, respectively). To accomplish this task, we adopted Gibbs sampling (Griffiths and Steyvers, 2004), a general method applied for probabilistic inference when direct sampling is difficult, which iteratively analyzes the set of training documents to estimate the probability of a word $w$ given a (latent) topic $t$ ($t$ given a document, respectively). The

11www.elkhart.k12.in.us/3_staff/curric/pdf/1eng.pdf. Note that even though we consider the single standard of expectations from the Elkhart (IN) public schools, the developers of TRoLL can freely extract different standards of expectations from states other than the state of Indiana, which will not change the design methodology of TRoLL in terms of using subject areas as a feature for predicting the readability level of a book.

12http://jgibblda.sourceforge.net/

13The training documents are uniformly distributed among the 55 pre-defined subject areas, i.e., 100 documents per subject area, and were retrieved by using a keyword query on each subject area $SA$ on Wikipedia so that the top-ranked retrieved Wikipedia page $P_{SA}$, along with the pages linked from $P_{SA}$, are treated as documents related to $SA$. 

8
Figure 3: Determining the subject area and grade level of a book \( Bk \) using its content

As shown in Figure 3, given an excerpt of \( Bk \), denoted \( Bk_e \), TRoLL uses the trained LDA model and Equation 2 to identify the (latent) topic covered in \( Bk_e \). Each potential latent topic of \( Bk_e \) is associated with a probability value which indicates its likelihood in describing \( Bk_e \). Thereafter, the topic \( T \) with the highest probability with respect to \( Bk_e \) is treated as the latent topic of \( Bk_e \).

\[
\text{Topic}(BK_e) = \arg\max_{T \in LT} P(T|BK_e) = \arg\max_{T \in LT} \sum_{i=1}^{\|BK_e\|} P(w_i|T)
\]  

(2)

where \( LT \) is the set of fifty-five pre-defined latent topics considered by the trained LDA model, \( P(T|BK_e) \) is the probability of \( T \) given \( BK_e \), \( \|BK_e\| \) is the number of distinct non-stop, stemmed words in \( BK_e \), \( w_i \) is the \( i^{th} \) word in \( BK_e \), and \( P(w_i|T) \) is the probability of \( w_i \) given \( T \) as determined by the trained LDA model.

Having identified \( T \) covered in \( BK_e \) using the trained LDA model, TRoLL applies Equation 3 to compute the subject area score (SAS) between \( T \) and each one of the fifty-five subject areas considered by TRoLL, which captures the degree of resemblance between (words in) \( T \) and (words in) the corresponding subject area. The subject area \( SA \) with the highest computed SAS is treated as the subject area of \( Bk \) based on its similarity with \( T \) and is assigned to \( Bk \). Hereafter, the grade level associated with \( SA \), which is determined by the mapping between US curriculum subject areas and grade levels employed by TRoLL (a portion of which is shown in Table 2), becomes the value of the content-based subject area predictor of \( Bk \).

\[
\text{Subject}(T) = \arg\max_{SA \in S} SAS(T, SA)
= \arg\max_{SA \in S} \frac{1}{|T|} \sum_{i=1}^{|T|} P(w_i|T) \times \frac{1}{|SA|} \sum_{j=1}^{|SA|} \text{wcf}(k_j, w_i)
\]

(3)

where \( S \) is the set of fifty-five subject areas, \( |T| \) (\( |SA| \), respectively) is the number of keywords in \( T \) (\( SA \), respectively), \( T, w_i \) and \( P(w_i|T) \) are as defined in Equation 2, \( k_j \) is the \( j^{th} \) word in \( SA \), \( \text{wcf}(k_j, w_i) \) is
Table 2: A sample of the fifty-five subject areas considered by TRoLL, along with their corresponding grades

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Grade</th>
<th>Subject Area</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapes</td>
<td>K</td>
<td>Geography</td>
<td>8</td>
</tr>
<tr>
<td>Addition</td>
<td>2</td>
<td>Mental disorders</td>
<td>9</td>
</tr>
<tr>
<td>Cultures</td>
<td>5</td>
<td>European history</td>
<td>11</td>
</tr>
<tr>
<td>Agriculture</td>
<td>6</td>
<td>Statistics</td>
<td>12</td>
</tr>
</tbody>
</table>

the word-correlation factor of $k_j$ and $w_i$ specified in the pre-defined word-correlation matrix (Koberstein and Ng, 2006), and $SA_n$ is the number of words in $SA$ that have a non-zero $wcf$ score with respect to words that define $T$.

Word-correlation factors in the correlation matrix, which is introduced in (Koberstein and Ng, 2006), reflect the degree of similarity between any two non-stop, stemmed words based on their (i) frequencies of co-occurrence and (ii) relative distances in a set of approximately 880,000 Wiki-pedia.org documents written by more than 89,000 authors that cover a wide variety of topics. Compared with synonyms/related words compiled by WordNet in which pairs of words are not assigned similarity weights, word-correlation factors offer a more sophisticated measure of word similarity.

Example 2 Consider the book “The scorpions of Zahir,” denoted $Bk_2$, written by Chris Brodien-Jones, which tells the story of a young girl who travels to the Moroccan desert with her family on a quest to save the ancient city of Zahir. Using Equations 2 and 3, TRoLL identifies “Cultures” as the subject area of $Bk_2$, which is taught in the 5th grade (see Table 2). Consequently, “5” is the value of the subject area predictor of $Bk_2$, which correlates with the publisher’s grade level range for $Bk_2$, which is 5 and up.

Analyzing Topical Information Metadata

In this section, we discuss the analysis of the metadata of a book $Bk$ based on its topical information, which are subject headings assigned to $Bk$ by professional catalogers who are certified by the Library of Congress or other book cataloging organizations. A subject heading is a set of keywords used by librarians to categorize and index books according to their themes. Subject headings take on several forms (Miller, 2006), which include the inverted form, e.g., “Trolls, Green,” the natural language form, e.g., “Green Trolls,” and the subdivision form, e.g., “Fantasy—Mythical Creatures—Trolls—Green.” Each component in a subdivision form is treated as a subject heading, whereas subject headings in inverted and natural language forms are each treated as a single subject heading. We discuss the predictors derived from subject headings of $Bk$ and the ones derived using the subject headings of books written by the author of $Bk$ below.

Book Subject Heading Predictors

To compute the predictors derived from the subject headings of a book $Bk$, TRoLL examines (i) their total count, (ii) their associated grade levels, and (iii) their rate of occurrence.

- **Total Count of Subject Headings.** TRoLL uses the count of subject headings assigned to $Bk$ as a predictor in Equation 1, since books that are more difficult to comprehend are often assigned more subject headings. We have empirically verified this claim by counting the number of subject headings assigned to each one of the 5,718 randomly chosen books (available at ARbookfind.com) with

[14]Wordnet.princeton.edu
Subject Headings and Grade Levels. Besides using the count of subject headings, TRoLL considers the subject headings of Bk that are previously encountered in books with a known readability level (range) recommended by their respective publishers. A previously encountered subject heading is a heading observed during the one-time mapping process of TRoLL, which paired subject headings assigned to each of the 8,737 books in the BookRL-SH dataset (introduced in the Experimental Results section) with the readability level range of the corresponding book determined by its publisher. To account for the possibility that a subject heading, $SH$, is paired with many books and therefore many readability levels, TRoLL considers all readability levels paired with $SH$ as a frequency distribution, $D$. An analysis of the mean, median, lower bound, and upper bound readability levels in $D$ yield four predictors, which are called frequency distribution predictors ($FDP$). Additionally, in order to reduce the effect of outlier readability levels in $D$, TRoLL further considers the mean, median, lower bound, and upper bound of the readability levels within one standard
deviation of the mean of \( D \), which generate another four predictors based on the mapping between subject headings and grade levels. The value of each of these eight predictors is calculated as

\[
FDP_{m_{i}}(Bk) = \frac{\sum_{j=1}^{|V|} m_{i}(D_{j})}{|V|}
\]

where \( m_{i} \) is either the mean, median, lowerBound, upperBound, SD\_mean, SD\_median, SD\_lowerBound, or SD\_upperBound, \( V \) is the set of all the subject headings assigned to \( Bk \) that have been previously encountered, \( D_{j} \) is the frequency distribution corresponding to a subject heading \( SH_{j} \in V \), and \( m_{i}(D_{j}) \) is the application of \( m_{i} \) to \( D_{j} \).

**Example 4** To illustrate how the mean frequency distribution predictor is calculated, let’s consider the three subject headings, i.e., \{aardvark, fiction, juvenile fiction\} = \( V \) (out of the five total), assigned to \( Bk_{3} \) (in Example 3) which have been previously encountered. According to Equation 4

\[
FDP_{\text{mean}} = \frac{\text{mean}(D_{\text{aardvark}}) + \text{mean}(D_{\text{fiction}}) + \text{mean}(D_{\text{juvenile fiction}})}{3}
\]

We observe that nine of the books used in the one-time mapping process described above were assigned the subject heading “aardvark”. The mean of the readability levels of the corresponding nine books, which are established by their publishers, are \( D_{\text{aardvark}} = <0, 0, 1.5, 1.6, 2.2, 2.3, 2.5> \). Based on this distribution, \( \text{mean}(D_{\text{aardvark}}) = 1.3 \). In the same manner, TRoLL examines the readability level distribution for “fiction” and “juvenile fiction” to compute \( \text{mean}(D_{\text{fiction}}) = 3.81 \) and \( \text{mean}(D_{\text{juvenile fiction}}) = 3.10 \). Subsequently, \( FDP_{\text{mean}} = \frac{1.3 + 3.81 + 3.10}{3} = 2.74 \), is the value of one of eight frequency distribution predictors based on the mean metric. □

- **Common Subject Headings.** Besides considering the mapping of subject headings to their grade levels, TRoLL also counts commonly occurred subject headings of \( Bk \). If a subject heading was previously encountered during the mapping process when 38,315 subject headings (assigned to the books in \( \text{BookRL-SH} \)) were examined, it is considered a commonly occurred subject heading. We conjecture that commonly occurred subject headings are assigned to books with lower readability levels, since books for lower readability levels cover less advanced, specific topics. The predictors created by using commonly occurred subject headings are (i) the number of previously encountered subject headings assigned to \( Bk \) and (ii) the ratio of previously encountered subject headings to the total number of subject headings assigned to \( Bk \).

**Example 5** Consider \( Bk_{3} \) in Example 3. Since the subject headings “aardvark,” “fiction,” and “juvenile fiction” have been previously encountered, whereas the others have not, 3 and \( \frac{4}{5} \) are the values of the number of previously encountered subject headings and the ratio of previously encountered subject headings predictors, respectively. □

**Author’s Subject Headings Predictors**

The subject headings assigned to books (including \( Bk \)) written by \( A_{Bk} \), who is the author of \( Bk \), are analyzed in the same manner as the subject headings assigned to \( Bk \). The analysis of commonly occurred subject headings assigned to books written by \( A_{Bk} \) is captured in one predictor, which is the ratio of the number of previously encountered subject headings to the total number of subject headings assigned to books written by \( A_{Bk} \). \( FDP_{\text{median}} \), which is based on the subject headings assigned to all the books written by \( A_{Bk} \), is established as another predictor. The median readability level was employed, since
medians are less influenced by outliers, which often decrease the accuracy of a frequency distribution predictor. Note that only the median, instead of all eight of the frequency distribution predictors is considered for $A_{Bk}$, since subject headings assigned to books written by $A_{Bk}$ are not always directly related to $Bk$, even though $A_{Bk}$ often writes books at a particular readability level.

**Example 6** Consider Stephen Krensky, the author of $Bk_3$ in Example 3. The books written by the author have been assigned fifty subject headings, which are shown in Figure 6. The ratio of previously encountered subject headings of books written by Krensky, $\frac{30}{50}$, is the value of the predictor for the author based on the previously encountered subject headings, whereas $FDP_{median}$ for Krensky, another TRoLL predictor, is calculated to be 2.6. □

**Analyzing Targeted Audience Metadata**

TRoLL also considers the audiences targeted by books and their corresponding authors in predicting the readability level of books.

**The Book Audience Level Predictor**

For each book in its database, OCLC provides an audience level, which is a numerical value between 0 and 1 that indicates “the type of reader believed to be interested in a particular book” and is publicly available at OCLC.\(^{15}\) We have observed that there is a correlation, which is not a direct relationship, between the audience level of a book $Bk$ and its readability level, which is expected, since authors often write at the reading comprehension level of their respective audiences (Crowhurst and Piche, 1979). The audience level of $Bk$ is the value of the book audience level predictor used by TRoLL.

**Example 7** Consider $Bk_3$ in Example 3 again. As depicted in the OCLC record for $Bk_3$ and shown in Figure 5, $Bk_3$ is aimed towards primary school readers with its audience level score being 0.1, which is the value of the corresponding book audience level predictor as specified in the mapping between targeted audiences and audience levels provided by OCLC and as shown in Figure 7. □

**The Author’s Audience Level Predictor**

Besides the audience level of $Bk$, OCLC also defines the audience level of its author $A_{Bk}$ as the average of the audience levels of the books written by $A_{Bk}$, including $Bk$. In addition, OCLC provides the minimum (maximum, respectively) audience level of books $A_{Bk}$ has written. Based on these three audience level scores determined by OCLC, we define three other audience level predictors: the average, lowest

\(^{15}\)http://www.oclc.org/research/activities/audience.html
Figure 7: The OCLC mapping between the targeted readers and their corresponding audience levels is available at http://www.oclc.org/research/activities/audience.html

Figure 8: The OCLC audience level for the author Stephen Krensky, available at http://www.worldcat.org/wcidentities/lccn-n79-109188

(minimum), and highest (maximum) audience levels of $A_{Bk}$, which refer to the comprehensive levels of the audience targeted by books written by $A_{Bk}$.

**Example 8** Consider Stephen Krensky who is the author of $Bk_3$ in Example 3. As depicted in the audience level record in OCLC and shown in Figure 8, the average, lowest, and highest audience levels for Stephen Krensky are 0.11, 0.10, and 0.15, respectively, which are the values of the corresponding three audience level predictors.

As illustrated in Figure 8, the audience level does not directly matches the grade level of an author. Instead, the audience level simply reflects the groups of readers targeted by an author at various levels (from values of 0 for kids to 1 for advanced readers).

**The Predicted Readability Level of a Book**

It is possible that some of the fifty-four predictors defined in Equation 1 for predicting the readability level of a book $Bk$ cannot be calculated, since their corresponding metadata or content may be missing. Hence, TRoLL defines a number of regression models, which are the variances of the one shown in Equation 1, that analyze diverse combinations of available predictors. Based on the distinct subsets of predictors that can be applied to books in the BookRL-RA dataset, there are 107 trained regression models used by TRoLL for predicting the readability levels of books.

With the calculated value of each of the predictors pertinent to $Bk$, TRoLL selects, among the trained regression models, the optimal one that includes the most (values of) predictors available for $Bk$ and that excludes any predictor not applicable to $Bk$ to compute the readability level of $Bk$.

**Example 9** Based on the information available online for $Bk_1$ as presented in Example 1, 52 predictors are applicable to $Bk_1$. Using the optimal regression model for $Bk_1$, the grade level of $Bk_1$ predicted by TRoLL is 6.8, which falls within the grade-level range, i.e., 5 to 7, defined for the book by its publisher. TRoLL also examines the 23 predictors applicable to $Bk_3$ as presented in Example 3 and predicts 0.98 as
Table 3: Sources of books used for creating BookRL

<table>
<thead>
<tr>
<th>Online Sources</th>
<th>Number of Books</th>
<th>Online Sources</th>
<th>Number of Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARbookfind</td>
<td>4,037</td>
<td>Penguin</td>
<td>600</td>
</tr>
<tr>
<td>Bookadventure</td>
<td>1,017</td>
<td>Simon &amp; Schuster</td>
<td>388</td>
</tr>
<tr>
<td>CLCD</td>
<td>6,667</td>
<td>YABC</td>
<td>3,038</td>
</tr>
<tr>
<td>Lexile</td>
<td>2,154</td>
<td>Yalsa</td>
<td>226</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>18,127</strong></td>
<td><strong>Total</strong></td>
<td><strong>18,127</strong></td>
</tr>
</tbody>
</table>

the readability level for \(Bk_3\) using the corresponding optimal regression model for \(Bk_3\), which correlates with the readability level, i.e., 1.0, defined by the publisher of \(Bk_3\). □

### Experimental Results

In this section, we first introduce the dataset and metric used for assessing the performance of TRoLL. Thereafter, we present the results of the empirical studies conducted for evaluating the effectiveness of TRoLL in grade level prediction and compare its prediction accuracy with existing widely-used readability formulas/analysis tools.

#### The Dataset

To the best of our knowledge, there is no existing benchmark dataset that can be used for assessing the performance of readability-level prediction tools on books. For this reason, we constructed our own dataset, BookRL, using data extracted from CLCD.com, a website established to assist teachers, parents, and librarians in choosing books for K-12 readers, Young Adults Book Central (Yabookscentral.com), Young Adults Library Service Association (ala.org/yalsa), ARbookfind.com, Lexile.com, and reputable publishers’ websites. (See Table 3 for the source websites and their numbers of books included in BookRL.) BookRL consists of 18,127 books distributed among the K-12 grade levels with their ranges determined by their publishers. Due to the lack of common consensus among researchers on the most accurate existing readability prediction tool (Benjamin, 2012), we consider publisher-provided grade levels as the “gold-standard,” since they are defined by human experts.

It is an easier task for a publisher to provide a range of grade levels for a book than a single readability level, since the latter requires precision, whereas the former an intelligent estimate. These human-assessed ranges of readability levels of books are adopted as the gold standard, which is applied to assess the performance of TRoLL and the readability formulas/analysis tools considered in our empirical study.

Among the 18,127 books in BookRL, a subset of 7,142 books, denoted BookRL-RA, was utilized to train the regression analysis model of TRoLL. Another subset of 8,737 books, denoted BookRL-SH, was employed by TRoLL to perform a one-time mapping between subject headings and readability levels, and the remaining subset of 2,248 books, denoted BookRL-Test, was used for assessing the performance of TRoLL and a number of well-known readability formulas/analysis tools. All the subsets of BookRL are disjoint.

#### Metrics

To assess the performance of TRoLL and other widely-used readability formulas/analysis tools, we compute their Mean Absolute Error (MAE) (Croft et al., 2010), each of which is the averaged absolute difer-
Figure 9: An analysis of the performance of (group of) predictors considered by TRoLL based on each book $B_k$ and its author $A_{B_k}$ in BookRL-Test.

Errors between the expected and the predicted grade levels of the books in BookRL-Test determined by the corresponding formula/tool.

$$MAE = \frac{1}{|BookRL-Test|} \sum_{B \in BookRL-Test} |PR(B) - GL(B)|$$  \hspace{1cm} (5)

where $|BookRL-Test|$ is the number of books in BookRL-Test, $GL(B)$ is the grade level of a book $B$ in BookRL-Test predicted by a readability formula/analysis tool, and $PR(B)$ is either the lower or upper bound of the grade level range of $B$ determined by its publisher, whichever is closest to $GL(B)$, which reflects the closeness of the predicted grade level to the grade level range of $B$.

We have also applied the Wilcoxon signed-rank test, which is a non-parametric test based on the differences between pairwise samples (Croft et al., 2010), to determine the statistical significance of the MAE on grade-level prediction obtained by TRoLL with respect to their counterparts obtained by various readability formulas/analysis tools.

**Performance Evaluation**

In this section, we (i) analyze the prediction accuracy of (various groups of) TRoLL’s predictors, (ii) verify the correctness of using content and/or metadata for readability-level prediction, and (iii) compare the performance of TRoLL with other readability analysis formulas/tools.

**Analyzing TRoLL’s Predictors**

TRoLL uses up to fifty-four predictors to determine the readability level of a book. As shown in Figure 1, these predictors can be grouped into seven categories according to the type of information on books and authors considered, which include traditional readability features, grammar concepts, subject areas, subject headings, and audience levels. Figure 9 shows the MAE obtained by each of the groups of predictors with the fraction of books in BookRL-Test to which the corresponding group of predictors is applicable. Based on the compiled results, we draw the following observations:

- The predictor on the audience level of a book provided by OCLC achieves the lowest MAE in readability-level prediction. This is anticipated, since there is a high correlation between the read-
ability level of a book and its targeted audience, even though there is no direct mapping between an 
audience level and a readability level. Unfortunately, the OCLC’s audience level for a book is not 
always available. For example, only 53% of the books in BookRL-Test are assigned an audience 
level. The same applies to the audience level of an author provided by OCLC, from where only 
66% of book authors in BookRL-Test are assigned an audience level.

• The subject area predictor receives the highest MAE, since books for emergent (K-3) readers tend 
to include more pictures than text and these non-textual contents are not utilized by TRoLL to iden-
tify US curriculum subject areas covered in books. However, this predictor is a suitable indicator 
of the readability levels of books targeting more advanced readers. Using BookRL-Test, we have 
empirically verified that this predictor yields at most a 0.14 MAE in analyzing the readability levels 
of books in the 5th to 8th grade levels.

• The most reliable predictors, which do not only achieve relatively low MAE but also are widely 
applicable, are the two groups that analyze subject headings. These groups of predictors rely on 
information frequently available for books and thus are applicable to the majority of books examined 
by TRoLL. As shown in Figure 9, predictors based on subject headings are applicable to at least 
97% of the books in BookRL-Test.

• The group of predictors based on traditional readability features and grammar concepts are effec-
tive; however, these predictors are computed on excerpts of books, which are seldom available. For 
example, only 6% of the books in BookRL-Test come with their corresponding excerpts.

Validating the Accuracy of Using Content, Topical Information, and Targeted Audiences in Predicting 
Readability Levels

TRoLL examines two major types of information to determine the readability levels of books: content (if 
it is available) and metadata of books. We have validated the prediction accuracy of TRoLL when distinct 
set of predictors based on content and/or metadata are considered using BookRL-Test.

• Using content-based information. The low MAE, which is 0.53, achieved by considering only 
the content-based predictors (as shown in Figure 10) is anticipated, since book content is a reliable 
source of information which has direct impact on the degree of difficulty in understanding the content 
of a book, even if only an excerpt of the book is available for analysis. The MAE obtained by using 
content-based predictors is based on the 127 books with content in BookRL-Test.

• Relying on information other than content. We have further observed that in estimating the readabil-
ity levels of books for emergent readers, relying solely on content can generate readability levels that 
do not correlate with the ones recommended by publishers of the corresponding books. For example, 
the MAE generated by using content-based predictors for 1st grade books in BookRL-Test with 
available sample text is 2.10, which is three times higher than the MAE (i.e., 0.70) generated using 
up to fifty-four predictors of TRoLL on books with sample content as shown in Figure 11. Real-
izing that considering only content information can lead to imprecisely-predicted readability levels 
of books for emergent readers, a fact that correlates with the study discussed in (Devk, 2006),16 
we have designed TRoLL so that it analyzes metadata on books with or without excerpts available 
online. In doing so, the MAE obtained using content- and metadata-based predictors on 1st grade 
books in BookRL-Test with available sample text decreases from the 2.1 (obtained solely based 
on content predictors) to 0.73.

16The study verifies that using contents of books for young readers to predict their readability levels tends to yield overstated 
readability levels for the books.
Using metadata. TRoLL considers two types of metadata predictors: topical information, i.e., subject headings, and targeted audience. The MAE obtained by using topical information predictors, which is 0.83 (as shown in Figure 10), is higher than the 0.75 overall MAE of TRoLL (as shown in Figure 11) but slightly lower than the MAE achieved by using only audience level predictors, which is 0.85 (as shown in Figure 10). This is expected, since subject headings are often available for books and is a consistent contributing factor in predicting the readability level of books, whereas audience levels are limited as opposed to other metadata/content predictors.

The overall performance of TRoLL. Based on the results of our conducted empirical study, we conclude that the readability prediction accuracy of TRoLL is consistent, regardless of the presence or absence of sample text of books. On the 127 books with available sample content in BookRL-Test, TRoLL achieves a 0.70 MAE (as shown in Figure 11), whereas among the 2,121 (= 2,248 - 127) books in BookRL-Test without sample text, the 0.76 MAE generated by TRoLL is within one grade level off the ranges specified by the publishers of the examined books. Moreover, the overall MAE of TRoLL on BookRL-Test, in which 94% of the (2,248) books are without text, is 0.75, which is only \( \frac{3}{4} \) of a grade level from the targeted grade level. This low MAE is not only an accomplishment of TRoLL, but also it cannot be achieved by any of the existing readability formulas/analysis tools, since none of them can predict the grade level of books without excerpts.
Table 4: Popular readability formulas employed in our empirical study

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coleman-Liau</td>
<td>[(0.0588 \times \text{Average number of letters per 100 words}) - (0.296 \times \text{Average number of sentences per 100 words}) - 15.8]</td>
</tr>
<tr>
<td>Flesch-Kincaid</td>
<td>[(0.39 \times \frac{\text{Number of words}}{\text{Number of Sentences}}) + (11.8 \times \frac{\text{Number of syllables}}{\text{Number of words}}) - 15.59]</td>
</tr>
<tr>
<td>Rix</td>
<td>[(0.121 \times \text{Average sentence length}) + (0.082 \times \text{Number of unique unfamiliar words}) + 0.659, \text{where unfamiliar words can be found at} \text{readabilityformulas.com/articles/spache-formula-word-list.php} ]</td>
</tr>
<tr>
<td>Spache</td>
<td>[(0.121 \times \text{Average sentence length}) + (0.082 \times \text{Number of unique unfamiliar words}) + 0.659, \text{where unfamiliar words can be found at} \text{readabilityformulas.com/articles/spache-formula-word-list.php} ]</td>
</tr>
</tbody>
</table>

Figure 12: Performance evaluation on TRoLL and other readability formulas based on the 127 books with excerpts in BookRL-Test

**Comparing TRoLL with Others**

Using the 127 (out of 2,248) books in BookRL-Test with excerpts, we compared the grade-level prediction accuracy of TRoLL with a number of well-known readability formulas based on text content: Coleman-Liau (Coleman, 1975), Flesch-Kincaid (Kincaid et al., 1975), Rix (Index) (Anderson, 1983), and Spache (Spache, 1953), which we have implemented based on their formulas that are shown in Table 4. (See discussion on these readability formulas/tools in Benjamin, 2012.)

Figure 12 shows that (i) the MAE of the grade level predicted by TRoLL for a book with text, which is 0.70 and is the same MAE shown in Figure 11, is slightly more than half of a grade from the grade (range) determined by its publisher and (ii) the MAE of TRoLL is at least 26% lower than the MAE created by its counterparts. The difference in MAE achieved by TRoLL over each of its counterparts is statistically significant, as determined using a Wilcoxon signed-ranked test with \( p < 0.001 \).

We have further compared the performance of TRoLL with two other popular readability analysis tools widely-used by grade schools and reading programs in the USA, the Accelerated Reader (AR) and Lexile. Even though the algorithms of AR and Lexile are not publicly accessible, we were able to find 897 books with AR scores and 314 books with Lexile scores among the books in BookRL-Test from ARbookfind.com and Lexile.com, respectively. As shown in Figure 13, TRoLL outperforms AR and is more accurate than Lexile in predicting the grade level of the analyzed books. The improvement in MAE achieved by TRoLL over either Lexile or AR is statistically significant as determined using a Wilcoxon signed-ranked test with \( p < 0.001 \).
Human Assessment on TRoLL

We further evaluated TRoLL to determine whether its predicted readability levels are perceived as accurate by ordinary users, which offers another perspective on the performance of TRoLL. The additional evaluation is based on real users’ assessments of TRoLL which goes beyond the performance analysis conducted and presented in previous subsections. To accomplish this task, we conducted a user study using Amazon’s Mechanical Turk,\footnote{https://www.mturk.com/mturk/welcome}, a “marketplace for work that requires human intelligence”, which allows individuals or businesses to programmatically access thousands of diverse, on-demand workers and has been used in the past to collect user feedback for multiple information retrieval tasks (Koolen et al., 2012).

In the user study, we considered a set of 10 sample books with diverse readability levels. (The list of books used in the study, along with their corresponding readability levels predicted by TRoLL, is shown in Table 5.) We created a HIT (Human Intelligent Task) on Mechanical Turk so that for each sample book $SB$, each appraiser was presented six different readability levels for $SB$ and asked to select the one that “best” captures the readability level of $SB$. The six readability levels were generated by AR, Coleman-Liau, Flesch-Kincaid, Rix, Spache, and TRoLL, respectively.

The user study was conducted between October 25 and October 30, 2013 on Mechanical Turk. Altogether, there were 127 responses among the HITs used in the study. Based on the corresponding set of responses provided by Mechanical Turk appraisers, we have verified that users tend to favor TRoLL’s predicted readability level for a given book. (The distribution of the 127 collected responses among the different readability-level prediction formulas/tools is shown in Figure 14.) Note that the larger number of users who favor TRoLL over the remaining readability formulas/tools is statistically significant, as determined using the Wilcoxon signed-ranked test with $p < 0.05$ for Flesch-Kincaid, Coleman-Liau, Rix,
Figure 14: Distribution of Mechanical Turk appraisers’ responses in choosing the reading levels of 10 books computed by various readability-level prediction formulas/tools

Figure 15: A screenshot of the online version of our readability prediction tool, TRoLL, which shows the readability level of a book, given its isbn number and Spache, and \( p < 0.001 \) for AR.

**Trollie, an Online Prototype of TRoLL**

We have implemented TRoLL and made it available as an online application, called Trollie. Through its user-interface, a user can either enter (a portion of) the title and/or author or isbn of a book, which is a unique identifier of the book. In the latter case, Trollie computes and presents the readability level of the corresponding book (through TRoLL, the back-end readability analysis tool) to the user. (See Figure 15 for an example.) In the former case, Trollie first conducts a search\(^{18}\) of books that match the keywords captured in the (portion of the) title and/or author provided by the user. Thereafter, if the title and/or author is not unique, i.e., if multiple books partially match the user-provided keywords, the user is required to select,\(^ {19} \) among the retrieved books, the desired one so that Trollie can generate its corresponding readability level. (See the screenshot of Trollie shown in Figure 16 for an example.)

By developing Trollie, which can be accessed either through its website (http://troll.cs.byu.edu/) or API (http://troll.cs.byu.edu/api), we facilitate the task of automatically determining the readability levels of books, which assists children and teenagers (parents and teachers, respectively) in locating books that they (their K-12 readers, respectively) can comprehend.

\(^{18}\)The search is currently powered by OpenLibrary.org.

\(^{19}\)To speed up its processing time, Trollie archives the readability levels of books that have been computed over time through its online interface. Thus, the previously-computed readability level of a book is instantly displayed; otherwise, Trollie computes the readability level of a book on-the-fly, whenever the calculate button is hit by the user. (See Figure 16 for an example.)
Conclusions and Future Work

Statistical data compiled over the last few years has shown that the reading ability of school-age children in America is falling in comparing with most of the developed countries in the world. It is essential to encourage children/teenagers to develop good reading habits, which is crucial for them to succeed at school and in the living of a good life, the mission statement of TRoLL, a tool for regression analysis of literacy levels developed by us.

TRoLL is unique compared with existing readability formulas/analysis tools, since it can predict the grade level of a book even without a sample text of the book by simply analyzing metadata on the book that is publicly accessible from popular online sources. Moreover, TRoLL considers the most commonly-used reading-level range, i.e., K-12 grade levels, to estimate the readability level of a book, as opposed to other reading scores or scales which are unintuitive and do not always mean anything for an individual. For example, it is much more appealing for an individual to know that the level of a book $B$ is “6” (on a K-12 scale) rather than “100”, since using the former it is easy to interpret that any reader at the sixth (or higher) grade level can understand $B$, which is not as easy by using the latter. Considering LCSH and subject areas, TRoLL analyzes the suitability of the content of a book in determining its readability level, which is significantly different from other existing approaches for predicting text readability by analyzing only shallow and/or linguistic features of a text. TRoLL is reliable, since it applies regression analysis on a number of predictors established by using textual features on books (if they are available), Library of Congress Subject Headings of books, US Curriculum subject areas identified in books, and information about book authors to predict the grade level of K-12 books. Unlike many of its counterparts, TRoLL can estimate the readability levels of books at different levels, i.e., from emergent to mature readers. Many existing readability formulas/tools (Begeny and Greene, 2014) are applicable to determine only the readability levels of text targeting either young or more mature readers, (ii) over- (under-) estimate the levels of a text, and/or (iii) are ineffective in determining the readability levels for emergent, i.e., young, readers.

The development of TRoLL is a significant contribution to the educational community, since grade levels predicted by TRoLL can be used by (i) teachers, parents, and school librarians to identify reading materials suitable to their K-12 readers and (ii) K-12 students as a guide in making their own reading
selections, which, in turn, can enrich their reading for learning experiences. Conducted empirical studies on TRoLL have verified not only its prediction accuracy, but also its superiority over existing readability formulas/analysis tools.

For future work, we plan to extend TRoLL so that it can be used for predicting the grade levels of reading materials other than books, such as articles posted on various websites, which should facilitate the process of locating different (educational) materials, besides books, that are suitable for K-12 readers.

References


