Beyond the Text: A Metadata-Based Readability Level Prediction Tool for K-12 Books

ABSTRACT

The readability level of a book is a useful measure for children, teenagers, teachers, parents, and school librarians to identify reading materials suitable for themselves or other K-12 readers. Unfortunately, relatively few published books are assigned a readability level by professionals, which 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 prevent book content from being made publicly accessible. To alleviate the text constraint imposed on readability analysis of books, we have developed TRoLL, which relies on metadata of books that is publicly and readily accessible from reputable book-affiliated online sources besides using textual features (if they are available) to predict the readability level of books. Based on a multi-dimensional analysis, TRoLL determines the grade level of any book instantly, even in the absence of sample text from the book, which is its uniqueness. Furthermore, TRoLL is a significant contribution to the educational community, since its computed readability levels of books 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 guarantee 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.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering, Selection Process

Keywords

Readability level analysis, book, K-12

WOODSTOCK '97 El Paso, Texas USA

1. INTRODUCTION

As reading is an essential skill, which can have significant impact on a youth's educational and future career development, 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 [25]. During the last century educators and researchers have dedicated resources to develop readability assessment tools/formulas which quantify the degree of difficulty of understanding a text [2, 12].

Traditional readability formulas, such as Flesch-Kincaid (Reading Ease) [19], 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[1]. 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 [12]. All of these formulas, however, require the existence of a (sample 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 [30] and Advantage-TASA Open Standard for Readability (ATOS) [28], two widely-used readability analysis tools specifically developed for analyzing the readability level of books.

To address the deficiencies of the design issues affecting existing readability formulas/analysis algorithms, we have developed a tool for regression analysis of literacy levels, denoted TRoLL, which considers metadata on 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 available online) to determine its subject area established by the US Curriculum, analyze various shallow features, and examine different grammatical concepts in Bk, (ii) the subject headings assigned to Bk, and (iii) the targeted audience of Bk.

TRoLL can be applied to predict the grade levels of K-12 books, which can serve as a guideline for young read-

[1]Davison and Kantor [9] claim that "nonsense text" can be deemed as easy-to-read by traditional readability formulas if it contains frequently-used, short words organized into brief sentences.

ers selecting books by themselves, 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" [21]. The grade levels predicted by TRoLL on textbooks and (non-)fictional books can also be used as a guidance for parents, teachers, and librarians in locating 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 of authors and books 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 [27].

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 to be analyzed [2]. 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 the corresponding publishers. Given the difficulty in obtaining large-scale datasets on books for training/testing a grade-level prediction tool on books [32], the constructed dataset is an asset to the research community.

The remainder of this paper is organized as follows. In Section 2, we discuss existing readability formulas/analysis tools. In Section 3, we detail the design methodology of TRoLL. In Section 4, we present the results of the empirical studies conducted to assess the design methodology of TRoLL and compare its performance with popular readability formulas/analysis tools. In Section 5, we provide a concluding remark and directions for future work.

2. RELATED WORK

For almost a century, formulas have been developed to determine the readability level or degree of difficulty of a text, resulting in hundreds of readability formulae/tools [33]. Traditional formulas, including Flesch-Kincaid [19] and Gunning Fog (Index) [16], are based on shallow features, only provide a rough estimation of the difficulty of a text, and are not always reliable [2, 12]. Lexile [30] and Advantage-TASA Open Standard for Readability (ATOS) [28], two well-known readability analysis tools, are based upon traditional features. While the former compares words in a text with 600 million words in the Lexile corpus to establish the semantic difficulty (word frequency) and syntactic complexity (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.

Besides the formulas/tools listed above, new approaches based on linguistic features have been developed [6, 14, 17, 29]. Coh-Metrix [14] 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, besides language and discourse characteristics. Collins-Thompson and Callan [6] rely on multiple statistical language models, which capture patterns of word usage in different grade levels, that are combined using the Naïve Bayes classification, to estimate the most probable grade level of a text. Schwarm and Ostendorf [29] 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. [17] consider lexical and grammatical features derived from syntactic structures to analyze the difficulty of T. (For a detailed discussion on commonly-used features for assessing the readability of a text, see [12].)

Qumsiyeh and Ng [26] and Ma et al. [22] have recently developed their own readability assessment tools. ReadAid [26] 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 [22] also considers visually-oriented features (such as the average font size and average 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 [26], ATOS [28], and SVM-Ranker [22], along with the aforementioned readability formulas, either partially or fully depend on the availability of at least a sample of the text to compute its grade level, which is a severe constraint as discussed earlier. TRoLL bypasses this constraint by using publicly available metadata on books to accomplish its task. (See [2, 33] for an in-depth discussion of other existing readability formulas.)

3. OUR READABILITY ANALYSIS TOOL

To overcome the reliance of existing readability formulas/analysis tools on the text of a book, and to improve upon one-dimensional approaches towards calculating readability levels of books, we introduce TRoLL, a sophisticated readability tool that can operate without book content (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[2], the Online Computer Library Center (OCLC)[3], and Open Library[4]. This freely accessible information often includes metadata, such as subject headings assigned to Bk, and occasionally includes the target audience and/or the partial/full text of Bk. The overall readability prediction process of TRoLL is depicted in Figure 1.

[2]http://www.loc.gov
[3]http://www.worldcat.org
[4]http://www.openlibrary.org

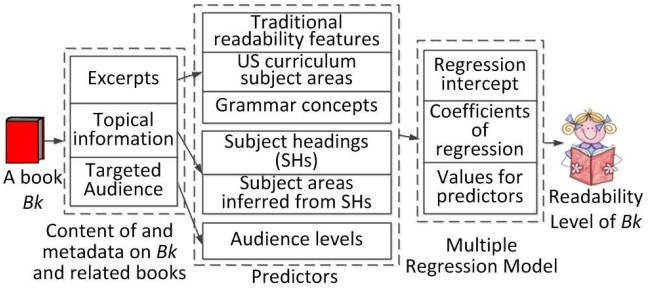

Figure 1: An overview of the readability level prediction process of TRoLL

3.1 Multiple Regression Analysis

To predict the readability level of a book Bk, TRoLL employs multiple linear regression analysis [34], which is a classical statistical technique for building estimation models. As shown in Equation 1, the model 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.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (1)$$

where y is the dependent variable, which is the predicted readability level of Bk, β_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 [34].

In Equation 1, each unknown parameter, i.e., the intercept parameter 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 [34] and the BookRL-RA training dataset (introduced in Section 4.1). Each book b in BookRL-RA is represented as a vector of the form $<b_1, b_2, \ldots, b_{57}, r>$, where b_i is the (value of the) the i^{th} predictor $(1 \leq i \leq 57)$ computed for book b, and r is the *target*, i.e., the known readability level for b in our case. (The fifty-seven predictors included in the regression model of TRoLL are explained in detail in Sections 3.2, 3.3, and 3.4, whereas the 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.

3.2 Analyzing Book Content

According to [4], only 7.7% of books in the OCLC database are linked to their partial or full content. We found similar results, since among the 7,142 books in the BookRL-RA dataset only 5% were linked to 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 [4]. 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) subject areas addressed in the book, or (iii) the grammar of its content. When calculating the values of these predictors, we only consider the first 2,500 characters[5] (to the nearest sentence) 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.

3.2.1 Predictors Based on Features Used by Traditional Readability Formulas

Existing widely-accepted readability formulas, such as Flesch-Kincaid [19], Coleman-Liau (Index) [5], Spache (Readability Index) [31], Gunning Fog [16], and SMOG (Index) [23], 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) syllables, (v) words with three or more syllables, (vi) unique unfamiliar words [31], and (vii) letters. 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 fourteen-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 that the target readers for the book should be in grades 5-7. Based on the first twenty pages of text of Bk_1 that are publicly available (a sample of which is shown in Figure 2), TRoLL analyzes a snippet of the first 2,639 characters (including the last sentence nearest to 2,500 characters) and calculates (the values of) the following predictors: Count of long words $= \frac{81}{2,639} = 0.031$, Count of sentences $= \frac{39}{2,639} = 0.015$, Count of total words $= \frac{475}{2639} = 0.032$, 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 $= \frac{86}{2,639} = 0.033$, and Count of letters $= \frac{2,018}{2,639} = 0.765$. □

3.2.2 The Subject Area Predictor on Book Content

TRoLL takes advantage of the mapping established by the US curriculum between subject areas and grade levels

[5]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 [13].

Wrapped in her *quilt*, Meg shook. She wasn't ***usually*** *afraid* of **weather**.--It's not just the **weather**, she **thought**.--It's the **weather** on top of **everything** else. On top of me. On top of *Meg Murry* doing **everything** wrong. School. School was all wrong. *She'd* been **dropped** down to the *lowest* **section** in her *grade*. That **morning** one of her **teachers** had said ***crossly***, "*Really*, *Meg*, I don't ***understand*** how a child with **parents** as ***brilliant*** as yours are ***supposed*** to be can be such a *poor* ***student***."

Figure 2: A sample of the text in "A Wrinkle in Time" in which long words are in bold, unfamiliar words are *italicized*, and words with three or more syllables are underlined

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 3^{rd} grade, whereas geometry at the 10^{th}. TRoLL pre-defines a number of subject areas to be considered, which is fifty-five. These subject areas were inferred from the K-12 curriculum posted under Elkhart Community School website[6].

To determine the subject area of a book Bk, TRoLL first analyzes (an excerpt of) its content by using a Latent Dirichlet Allocation (LDA) model [3], 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 [3]. To train a LDA model, we pre-defined the number of latent topics to be fifty-five (to match the number of subject areas considered by TRoLL) and applied JGibbLDA[7], a Java implementation of LDA, on 5,500 training documents randomly chosen from Wikipedia.org[8]. 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 [15], 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 sampling method is efficient and has been successfully used for obtaining good approximations for LDA [18].

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 potential (latent) topics covered in Bk_e. Each latent topic is associated with a probability value which indicates its likelihood in describing Bk_e. Thereafter, the topic T with the *highest* probability is treated as the *topic* of Bk.

$$\begin{aligned} Topic(BK_e) &= \text{argmax}_{T \in LT}\ P(T|BK_e) \\ &= \text{argmax}_{T \in LT} \sum_{i=1}^{|BK_e|} P(w_i|T) \qquad (2) \end{aligned}$$

[6]www.elkhart.k12.in.us/3_staff/curric/pdf/1eng.pdf

[7]http://jgibblda.sourceforge.net/

[8]The 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.

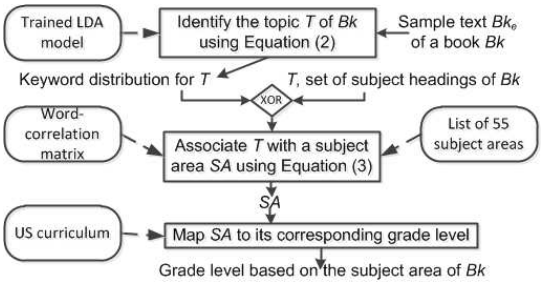

Figure 3: Determining the subject area and grade level of a book Bk using its content/subject headings

where LT is the set of fifty-five 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. Hereafter, the grade level associated with SA assigned by the US curriculum becomes the value of the *content-based subject area* predictor of Bk.

$$\begin{aligned} &Subject(T) \\ &= \text{argmax}_{SA \in S}\ SAS(T, SA) \\ &= \text{argmax}_{SA \in S} \frac{1}{|T|} \sum_{i=1}^{|T|} P(w_i|T) \times \frac{1}{SA_n} \sum_{j=1}^{|SA|} wcf(k_j, w_i) \quad (3) \end{aligned}$$

where SA is one of the fifty-five subject areas in S, $|T|$ ($|SA|$, respectively) is the number of keywords in T (SA, respectively), T and w_i and $P(w_i|T)$ are defined in Equation 2, k_j is the j^{th} word in SA, $wcf(k_j, w_i)$ is the *word-correlation factor* of k_j and w_i specified in the pre-defined word-correlation matrix [20], 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 introduced in [20] 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 Wikipedia.org documents written by more than 89,000 authors that cover a wide variety of topics. Compared with synonyms/related words compiled by WordNet[9] in which pairs of words are not assigned similarity weights, word-correlation factors offer a more sophisticated measure of word similarity.

3.2.3 Grammar Predictors on Book Content

TRoLL examines grammatical constructions, as defined by the US curriculum and shown in Table 1, to compute the value of grammar predictors. These predictors reflect the *complexity* of the (i) writing style, (ii) organization of the

[9]Wordnet.princeton.edu

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 using the Stanford NLP Parser [10], than the analysis used in Flesch-Kincaid [19], Coleman-Liau [5], SMOG [23], 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 *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 [10] to parse the text into *parse trees* (see Table 1 for the list of grammatical constructions analyzed using the Stanford Parser). 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. Even though the counts per simple grammatical concepts and counts per parse tree do not directly determine the readability level of a text, they correlate to its readability level and hence are useful in predicting the readability of Bk using Equation 1.

3.2.4 The Subject Area Predictor of an Author

TRoLL analyzes the content that describes the published works of the first author[10] of Bk, given that in general an author consistently writes books at a certain readability level.

TRoLL predicts the grade levels of authors' targeted readers based on the subject areas covered in books written by their corresponding authors. Using Equation 2, TRoLL identifies the topic with the highest probability in capturing the content of each book written by the author A of Bk. Hereafter, TRoLL applies Equation 3 to determine the subject areas paired with the detected topics. Finally, the grade levels of the identified subject areas are *averaged* to yield the value of the *author content-based subject area* predictor for Bk.

3.3 Analyzing Book Metadata–Topical Info.

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 topics. Subject headings take on several forms [24], including *inverted form*, e.g., "Trolls, Green," *natural language form*, e.g., "Green Trolls," and *subdivision form*, e.g., "Fantasy—Mythical Creatures—Trolls—Green." *Each* subdivision in a subdivision form is treated as a subject heading, whereas subject headings in inverted or natural language form 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.

[10]We have empirically verified that by considering only the *first* author of a 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 more than one author.

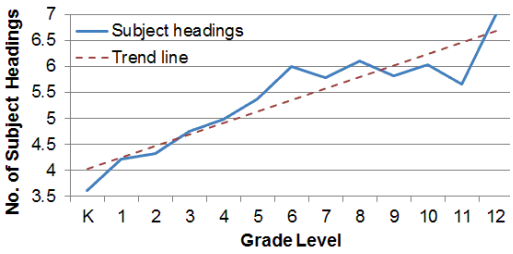

Figure 4: The relationship between the number of subject headings assigned to books and their readability levels determined by AR

3.3.1 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 its readability levels determined by Accelerated Reader (AR). The mapping between the number of subject headings and grade levels is depicted in Figure 4. The trend line in Figure 4 has a positive slope of about $\frac{2}{9}$, which demonstrates that books of high readability levels are assigned, on average, more subject headings than books of lower reading levels.

Example 2. Consider "Arthur and the Cootie Catcher," denoted Bk_2, which is a book written by Stephen Krensky and included in the *Arthur the Aardvark children's series*. Bk_2 was assigned *aardvark, cootie catchers, fiction, fortune telling*, and *juvenile fiction* as its subject headings. Thus, five (the number of subject headings) is the value of the count for the *Book Subject Heading* predictor of Bk_2, one of the predictors used in Equation 1 for predicting the readability level of Bk_2. □

3.3.2 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 a 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 Section 4.1) with the readability level of the corresponding book. 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 be-

Table 1: List of predictors used by TRoLL

Predictors Based on Content (38)			
Predictors Based on Traditional Text Features (7)			
Count of long words	Count of sentences	Count of total words	Count of letters
Count of syllables	Count of words with three or more syllables	Count of unique unfamiliar words	
Content-based subject area predictor (1)			
Predictors Based on Grammatical Constructions (29)			
Simple		Parse-tree Based	
Common prefixes (un-, re-, pre-, in-, de-, dis-)	Plural words	Adverbial phrases	Independent clauses
Conjunctions (and, but, or)	Personal pronouns (him, her, it)	Adverbs	Interrogative sentences
Conjunctive adverbs (however, therefore, on the other hand)	Possessive nouns	Comparatives and superlatives	Modal verbs of deduction
Contractions	Prepositions	Consecutive verbs	Participles
Determiners	Suffixes (-er, -ment, -able, -ness, -ly, -ful, -less, -tion, -ight, -ite, -ate)	Dependent clauses	Past progressive tense
Irregular vowel combinations, spelling, or phonetics (boot, soil, trout)	Syncategorematic words (like, as, to, if, all)	First conditional form	Past tense
		Future tense	Prepositional phrases
		Quantifiers	Present perfect tense
			Present progressive tense
Author content-based subject area predictor (1)			
Predictors Based on Topical Information (15)			
Total count of subject headings (1)			
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 (1)			
Ratio of previously encountered subject headings (1)			
Subject area predictor based on subject headings (1)			
Ratio of previously encountered subject headings assigned to books written by a given author (1)			
Median of readability levels paired with subject headings assigned to books written by an author (1)			
Author subject area predictor based on subject headings (1)			
Predictors based on Targeted Audience (4)			
Book audience level (1)		Average author audience level (1)	
Minimum author audience level (1)		Maximum author audience level (1)	

tween subject headings and grade levels. The value of each of these eight predictors is calculated as

$$FDP_{m_i}(Bk) = \frac{\sum_{j=1}^{|V|} m_i(D_j)}{|V|} \quad (4)$$

where m_i is either *mean, median, lowerBound, upperBound, SD_mean, SD_median, SD_lowerBound*, or *SD_upperBound*, V is the set of all 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 3. To illustrate how the *mean* frequency distribution predictor is calculated in practice, let's consider the three subject headings (out of the five total) assigned to Bk_2 (in Example 2) which have been previously encountered, i.e., $V = \{$*aardvark, fiction, juvenile fiction*$\}$. According to Equation 4

$$FDP_{mean} = \frac{mean(D_{aardvark}) + mean(D_{fiction}) + mean(D_{juvenile\ fiction})}{3}$$

We observe that *nine* of the books used in the one-time mapping process described above were assigned the subject heading *aardvark*. The corresponding readability levels of these nine books are $D_{aardvark}$ = <0.0, 0.0, 0.0, 1.5, 1.6, 1.6, 2.2, 2.3, 2.5>. Based on this distribution, $mean(D_{aardvark})$ = 1.3. In the same manner, TRoLL examines the readability levels distribution for *fiction* and *juvenile fiction* to compute $mean(D_{fiction})$ = 3.81 and $mean(D_{juvenile\ fiction})$ = 3.10. Subsequently, $FDP_{mean} = \frac{1.3+3.81+3.10}{3} = 2.74$, is the value of one of eight *frequency distribution predictors* based on the mean metric. □

3.3.3 Common Subject Headings

Besides mapping subject headings to their grade levels,

TRoLL also considers *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 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 4. Consider Bk_2 in Example 2. Since the subject headings *aardvark, fiction*, and *juvenile fiction* have been previously encountered, whereas the others have not, 3 and $\frac{3}{5}$ are the values of the two predictors associated with the *commonly occurred* subject headings, respectively. □

3.3.4 Subject Area Predictor Using Subject Headings

As previously stated, sample text for a book is often not publicly accessible online. For this reason, TRoLL explores other alternatives to determine the subject area of a book. Given that subject headings (i) are keywords assigned to books by the Library of Congress to topically categorize books and (ii) are publicly available, they can be treated as the topic addressed in a book Bk, instead of the latent topic identified by the trained LDA model based on the content of Bk (introduced in Section 3.2.2).

As shown in Figure 3, TRoLL treats the *subject headings* assigned to Bk as the set of keywords that define the topic T of Bk and applies Equation 3 to determine the subject area SA with the highest $SAS(T, SA)$ score (among all the subject areas) as the *subject area* of Bk. The grade level associated with SA is the value of the *subject area predictor* based on subject headings of Bk.

3.3.5 An Author's Subject Headings Predictors

The subject headings assigned to books written by an author A are analyzed in the same manner as the subject headings assigned to a book Bk (as discussed in Sections 3.3.2 and 3.3.3). The analysis of *commonly occurred* subject headings assigned to books written by A 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. FDP_{median}, which is based on the subject headings assigned to all the books written by A, 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 defined in Section 3.3.2 is considered for A, since subject headings assigned to books written by A are not always directly related to Bk, even though A often writes books at a particular readability level.

EXAMPLE 5. Consider Stephen Krensky, the author of Bk_2 in Example 2. The books written by the author have been assigned *fifty* subject headings, which are shown in Figure 5. 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 commonly occurred subject headings, whereas FDP_{median} for Krensky, another TRoLL predictor, is calculated to be 2.6. □

Figure 5: Subject headings assigned to books written by Stephen Krensky, available at http://www.worldcat.org/wcidentities/lccn-n79-109188

3.3.6 An Author's Subject Area Predictor based on Subject Headings

TRoLL also examines the subject area covered in each book written by the same author A of a book Bk by using subject headings assigned to books. As discussed in Section 3.3.4, the topic of each book authored by A is captured using the corresponding set of subject headings. Thereafter, the grade level of each subject area that most closely resembles the subject headings assigned to each book written by A (identified using Equation 3) is *averaged* to generate the value of the *author subject area predictor based on subject headings*.

3.4 Analyzing Book Metadata–Targeted Audience

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

3.4.1 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[11]. We have observed that there is a correlation 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 [8]. The audience level of Bk is the value of the *book audience level* predictor used by TRoLL.

3.4.2 The Author Audience Level Predictor

OCLC computes the *audience level* of an author A as the *average* of the audience levels of the books written by A. In addition, OCLC provides the *minimum* (*maximum*, respectively) audience level of books A has written.

Based on the three audience level scores determined by OCLC, we define three other audience level predictors: the *average, lowest*, and *highest* audience levels of A, which are based on the audience levels of the books written by A.

EXAMPLE 6. Consider Stephen Krensky who is the author of Bk_2 in Example 2. As depicted in the audience level record in OCLC and shown in Figure 6, 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. □

[11]http://www.oclc.org/research/activities/audience.html

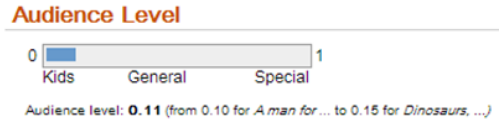

Figure 6: The OCLC audience level for the author Stephen Krensky, available at http://www.worldcat.org/wcidentities/lccn-n79-109188

3.5 The Predicted Readability Level of a Book

It is possible that some of the fifty-seven 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 considered by TRoLL for predicting the readability levels of books.

With the calculated values 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 without any predictor not applicable to Bk to compute the readability level of Bk.

EXAMPLE 7. Based on the information available online for Bk_1 as presented in Example 1, 55 predictors are applicable to Bk_1. Using the optimal regression model for Bk_1, the grade level of Bk_1 predicted by TRoLL is 7.7, which is close to the grade-level range, i.e., 5 to 7, defined for the book by its publisher. □

4. EXPERIMENTAL RESULTS

In this section, we first introduce the dataset and metric (in Sections 4.1 and 4.2, respectively) 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 performance with existing widely-used readability formulas/analysis tools (in Section 4.3).

4.1 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 (Yaboo.com), Young Adults Library Service Association (ala.org/yalsa), ARbookfind.com, Lexile.com, and reputable publishers' websites. (See Table 2 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 grade levels (ranges) determined by their publishers. Due to the lack of consensus among researchers on the most accurate existing readability prediction tool [2], we consider publisher-provided grade levels as the "gold-standard," since they are defined by human experts.

Among the 18,127 books in BookRL, a subset of 7,142 books, denoted BookRL-RA, was employed by TRoLL to perform its one-time mapping between subject headings and readability levels (as discussed in Section 3.3). Another subset of 8,737 books, denoted BookRL-SH, was utilized to train the regression analysis model of TRoLL (as introduced in Section 3.1), and the remaining 2,248 books, denoted BookRL-Test, were used for assessing the design methodology and prediction accuracy of TRoLL, in addition to comparing its performance with a number of well-known readability formulas/analysis tools. (Note that BookRL-RA, BookRL-SH, and BookRL-Test are disjoint.)

Table 2: Sources of books used for creating BookRL

Online Sources	No. of Books	Online Sources	No. of Books
ARbookfind	4,037	Penguin	600
Bookadventure	1,017	Simon & Schuster	388
CLCD	6,667	Yaboo	3,038
Lexile	2,154	Yalsa	226
		Total	18,127

4.2 Metrics

To assess the performance of TRoLL and other widely-used readability formulas/analysis tools, we compute their (absolute) error rates (ER) [7], each of which is the *absolute difference* between an *expected* and a *predicted* grade level for a book B.

$$ER = \frac{1}{|BookRL - Test|} \sum_{B \in BookRL-Test} |PR(B) - GL(B)| \quad (5)$$

where |BookRL-Test| is the number of books in BookRL-Test, $GL(B)$ is the grade level of B 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 and was applied to all the examined prediction tools/formulas.

4.3 Performance Evaluation

In this section, we verify the correctness of the design methodology of TRoLL and compare its performance with others (in Sections 4.3.1 and 4.3.2, respectively).

4.3.1 Evaluating the Design Methodology of TRoLL

As discussed in Section 3, TRoLL examines two major types of information to determine the readability levels of books: content and metadata of books. We validate the correctness of the design methodology of TRoLL by conducting an empirical study using the BookRL-Test dataset, which measures the prediction accuracy of TRoLL when distinct set of predictors based on content and/or metadata are considered.

In comparing the usage of content versus metadata predictors, the error rate of the former is lower than the latter as shown in Figure 7, which is not surprising, 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 error rate achieved by using the content predictors is based on the 127 books with content in BookRL-Test, as opposed to the 2,248 books used for eval-

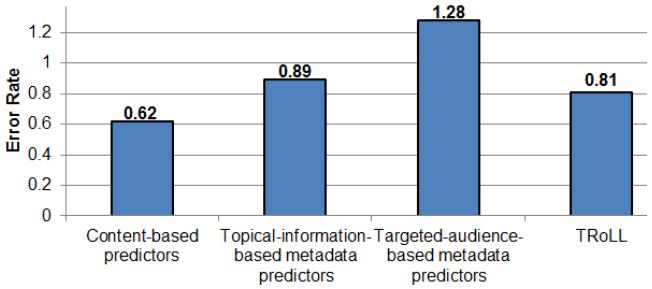

Figure 7: Performance evaluation of TRoLL using distinct sets of predictors based on book contents and metadata

uating content and/or metadata predictors.

Realizing that considering only content information can lead to imprecisely-predicted readability levels of books for emergent readers[12] [11], we have designed TRoLL so that it analyzes metadata on books that are either with or without excerpts available online. Further examination of different types of metadata predictors used by TRoLL reveal their capabilities in accurately predicting the readability level of books. As discussed in Sections 3.3 and 3.4, TRoLL considers two types of metadata predictors: topical information, i.e., subject headings, and targeted audience. The error rate obtained by using topical information predictors is *higher* than the overall error rate of TRoLL but *lower* than the error rate achieved by using audience levels predictors as shown in Figure 7. This is anticipated, since subject headings are often available for books and thus become a consistent contributing factor in predicting the readability level of books. The higher error rate imposed on the audience level predictors is expected, since these predictors are limited in number, as opposed to other metadata/content predictors.

4.3.2 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: Coleman-Liau [5], Flesch-Kincaid [19], Rix (Index) [1], and Spache [31] , which we have implemented. (See discussion on the readability formulas/tools in Section 2 and [2]). Figure 8 shows that (i) on average the grade level predicted by TRoLL for a book with text (in BookRL-Test) is about *half* of a grade from the grade (range) determined by its publisher and (ii) the error rate of TRoLL is *at least* 33% lower than the error rate created by its counterparts.

We further compare the performance of TRoLL with two popular readability analysis tools widely-accepted by grade schools and reading programs in the USA: Accelerated Reader (AR) and Lexile. Recall that the algorithms developed to compute AR and Lexile scores are not publicly accessible, but we were able to find 897 books with AR scores and 314 books with Lexile scores, among the books

[12]We have empirically verified that when estimating the readability levels of books for emergent (i.e., K-2) readers, relying solely on content can generate readability levels that do not correlate with the ones recommended by publishers of the corresponding books—the average error rate generated by using content-based predictors for 1^{st} grade books in BookRL-Test is 2.10, whereas the error rate obtained using the metadata-based predictors is 0.73.

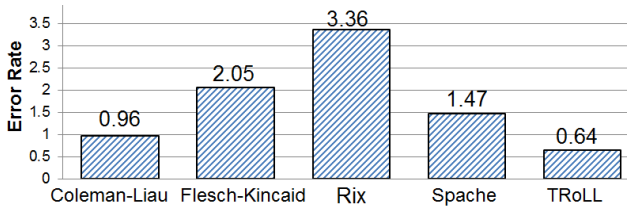

Figure 8: Performance evaluation on TRoLL and other readability formulas

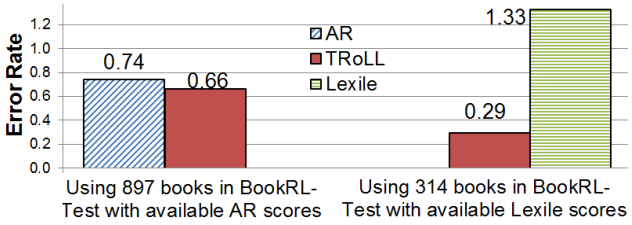

Figure 9: Performance evaluation on TRoLL, AR, and Lexile

in BookRL-Test, from ARbookfind.com and Lexile.com, respectively. As shown in Figure 9, TRoLL outperforms AR and is significantly more accurate than Lexile in predicting the grade level of the analyzed books (in BookRL-Test).

We have also evaluated the performance of TRoLL in predicting the grade level of books for which their excerpts cannot be obtained online. Among the 2,121 books in BookRL-Test without sample text, the 0.82 error rate generated by TRoLL is *less than one* grade level off the ranges specified by the publishers of the books. This low error rate is not only an impressive 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. The overall error rate of TRoLL on BookRL-Test, in which 94% of the books are without text, is 0.81 (see Figure 7), which is within one grade level of the targeted grade level.

5. CONCLUSIONS

We have introduced TRoLL, which 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. TRoLL is *novel* and *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.

Statistical data compiled over the last few years has shown that the reading ability of school-age children in America is falling in comparison to most of the developed countries in the world; it is essential to encourage children/teenagers to develop good reading habits, which is crucial to succeed at school and in the living of a good life, the mission statement of TRoLL. Grade levels predicted by TRoLL can be used by (i) teachers, school librarians, and parents 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. A conducted empirical study on TRoLL has 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 to predict 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 that are suitable for K-12 readers.

6. REFERENCES

[1] J. Anderson. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, 26:993–1022, 1983.

[2] R. Benjamin. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24:63–88, 2012.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003.

[4] X. Chen. Google Books and WorldCat: A Comparison of Their Content. *Online Information Review*, 36(4):507–516, 2012.

[5] M. Coleman. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975.

[6] K. Collins-Thompson and J. Callan. A Language Modeling Approach to Predicting Reading Difficulty. In *HLT-NAACL*, pages 193–200, 2004.

[7] W. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. Addison Wesley, 2010.

[8] M. Crowhurst and G. Piche. Audience and Mode of Discourse Effects on Syntactic Complexity in Writing at Two Grade Levels. *Research in the Teaching of English*, 13(2):101–109, 1979.

[9] A. Davison and R. Kantor. On the Failure of Readability Formulas to Define Readable Texts: A Case Study from Adaptations. *Reading Research Quarterly*, 17(2):187–209, 1982.

[10] M. De Marneffe. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC*, pages 449–454, 2006.

[11] ATOS vs. Lexile Which Readability Formula is Best? Renaissance Learning. kmnet.renlearn.com/Library/R003520002GE7114.pdf, 2006.

[12] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A Comparison of Features for Automatic Readability Assessment. In *COLING*, pages 276–284, 2010.

[13] D. Friedman and L. Hoffman-Goetz. A Systematic Review of Readability & Comprehension Instruments Used for Print and Web-Based Cancer Information. *Health Education & Behavior*, 33(3):352–373, 2006.

[14] A. Graesser, D. McNamara, M. Louwerse, and Z. Cai. Coh-Metrix: Analysis of Text on Cohesion and Language. Behavior Research Methods. *Instruments and Computers*, 36(2):193–202, 2004.

[15] T. Griffiths and M. Steyvers. Finding Scientific Topics. *PNAS*, 101:5228–5235, 2004.

[16] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[17] M. Heilman, K. Collins-Thompson, and M. Eskenazi. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *NAACL-HLT BEA8*, pages 71–79, 2008.

[18] Q. Jiang, J. Zhu, M. Sun, and E. Xing. Monte Carlo Methods for Maximum Margin Supervised Topic Models. In *NIPS*, pages 1601–1609, 2012.

[19] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease formula) for Navy Enlisted Personnel. Technical Report 8-75, Chief of Naval Technical Training, 1975.

[20] J. Koberstein and Y.-K. Ng. Using Word Clusters to Detect Similar Web Documents. In *KSEM*, pages 215–228, 2006.

[21] The Development of ATOS: The Renaissance Readability Formula. http://doc.renlearn.com/KMNet/R004250827GJ11C4.pdf, 2012.

[22] Y. Ma, E. Fosler-Lussier, and R. Lofthus. Ranking-based Readability Assessment for Early Primary Children's Literature. In *NAACL-HLT*, pages 548–552, 2012.

[23] G. McLaughlin. SMOG Grading–A New Readability Formula. *Reading*, 12(8):639–646, 1969.

[24] J. Miller. *Cataloging Correctly for Kids: An Introduction to the Tools, 4th Ed.*, chapter Sears list of subject Headings, pages 75–79. American Library Association, 2006.

[25] J. Oakhill and K. Cain. The Precursors of Reading Ability in Young Readers: Evidence from a Four-Year Longitudinal Study. *School Science Review*, 16(2):91–121, 2012.

[26] R. Qumsiyeh and Y.-K. Ng. ReadAid: A Robust and Fully-Automated Readability Assessment Tool. In *IEEE ICTAI*, pages 539–546, 2011.

[27] Matching Books to Students: How to Use Readability Formulas and Continuous Monitoring to Ensure Reading Success. http://doc.renlearn.com/KMNet/R003544312GE0BA6.pdf, 2011.

[28] I. School Renaissance Inst. The ATOS Readability Formula for Books and How it Compares to Other Formulas. Technical Report ED449468, ERIC Document Reproduction Service, 2000.

[29] S. Schwarm and M. Ostendorf. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *ACL*, pages 523–530, 2005.

[30] D. Smith, A. Stenner, I. Horabin, and M. Smith. The Lexile Scale in Theory and Practice: Final Report. Technical Report ED307577, ERIC Document Reproduction Service, 1989.

[31] G. Spache. A New Readability Formula for Primary-Grade Reading Materials. *Elementary School*, 53(7):410–413, 1953.

[32] K. Tanaka-Ishii, S. Tezuka, and H. Terada. Sorting Texts by Readability. *Computational Linguistics*, 36(2):203–227, 2010.

[33] The Principles of Readability. www.nald.ca/library/research/readab/readab.pdf, 2004.

[34] J. Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western Pub, 2009.