Statistical Estimation of Word Acquisition With Application to Readability Prediction

Paul KIDWELL, Guy LEBANON, and Kevyn COLLINS-THOMPSON

Models of language learning play a central role in a wide range of applications: from psycholinguistic theories of how people acquire new word knowledge, to information systems that can automatically match content to users' reading ability. Traditional methods for estimating word acquisition ages or content readability are typically based on linear regression over a small number of summary features derived from time-consuming user studies or costly expert judgments. With the increasing amounts of content available from the web and other sources, however, new statistical approaches are possible that can exploit this easily acquired data to learn more flexible, fine-grained models of language usage. We present a novel statistical model for document readability that is based on the logistic Rasch model and the quantiles of word acquisition age distributions. We use this model to estimate the distributions of word acquisition ages from empirical readability data collected from the web. We then demonstrate that the estimated acquisition distributions are very effective in predicting both global and local document readability. We also compare the estimated distributions with word acquisition data from existing oral studies, revealing interesting historical trends as well as differences between oral and written word acquisition grade levels.

KEY WORDS: Rasch model; Readability.

1. INTRODUCTION

Word acquisition refers to the temporal process by which children learn the meaning and understanding of new words. Some words are acquired at a very early age, some are acquired at early primary school grades, and some are acquired at high school or even later in life as the individual undergoes experiences related to that word. Different people acquire words at different ages and given a population of individuals $\mathcal{T}$, we refer to the acquisition age distribution of different words $w \in V$ resulting from random draws from the population $\mathcal{T}$

$$p_w(t) = P(\text{random person from } \mathcal{T} \text{ acquired } w \text{ at age } t). \quad (1)$$

In this paper we assume that the population $\mathcal{T}$ is infinite, and t is a continuous quantity. These relatively mild assumptions are made for convenience. Removing them adds to the complexity of the model and its presentation.

A related concept to acquisition age is document grade-level readability which refers to the school grade level of the document's intended audience. It applies in situations where documents are written with the expressed intent of being understood by children in a certain school grade. For example, textbooks or newsletters authored specifically for fourth graders are said to have readability grade level four or readability age 10. Defining readability formally requires some care as it implies that documents are readable by a population of individuals having acquired potentially different words. We define this concept formally, in agreement with previous studies, in the next section.

In this paper we explore a statistical model that draws a connection between document grade level readability and the acquisition age distributions (1). The model is a variation of the logistic Rasch model, based on the quantiles of the word acquisition distributions (1). We then demonstrate how the model may be used to estimate the word acquisition distributions p_w, $w \in V$, from document readability data collected by crawling the web, and then follow up with using the model for predicting the readability level of new documents. Such predictive abilities are highly relevant to web search technology as they can ensure that retrieved content is readable by a user who is searching the web. This can be done as follows: the user specifies the search query along with their age. Traditional information retrieval (Baeza-Yates and Ribeiro-Neto 1999) may be used to match content of websites to the query while the readability model may be used to filter out inappropriate readability levels.

In our experimental study we use three different datasets of documents annotated with readability grade level obtained from the internet. We examine the inferred acquisition distributions by analyzing and contrasting them with previous studies on oral word acquisition. The comparison between the estimated acquisition distributions and the previous studies reveals interesting historical trends in language learning as well as differences in the grade level of oral and written word acquisition. While some previous work has been done on this topic, our work is the first to draw a mathematical connection between readability and word acquisition and to use readability data for estimating the grade level of word acquisition. Interestingly, the proposed model performs very well in predicting readability of new documents in practice, achieving an error rate lower than alternative models.

2. A MODEL FOR DOCUMENT READABILITY AND WORD ACQUISITION

For a fixed word and a fixed population of individuals $\mathcal{T}$ the age of acquisition (AoA) distribution p_w represents the age

Paul Kidwell is Statistician, Lawrence Livermore National Laboratory, Livermore, CA 94550 (E-mail: *kidwellpaul@gmail.com*). Guy Lebanon is Assistant Professor, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332 (E-mail: *lebanon@cc.gatech.edu*). Kevyn Collins-Thompson is Researcher, Microsoft Research, Redmond, WA 98052 (E-mail: *kevynct@microsoft.com*). The authors thank Joshua Dillon for downloading the weekly reader data and preprocessing it. Constructive comments by the editor and reviewers helped improve the paper. The work described in this paper was funded in part by NSF grant DMS-0604486 and prepared in accordance with LLNL contract DE-AC52-07NA27344.

Journal of the American Statistical Association
March 2011, Vol. 106, No. 493, Applications and Case Studies
DOI: 10.1198/jasa.2010.ap09318

at which word w was acquired by the population [see Equation (1)]. In homogeneous populations p_w is likely to be unimodal, with most of its mass concentrated around a typical acquisition age μ_w. A reasonable choice for such a parametric family is the truncated normal distribution

$$p_w(t) \propto \mathrm{N}(t; \mu_w, \sigma_w) = (2\pi\sigma_w^2)^{-1/2} \exp(-(t-\mu_w)^2/(2\sigma_w^2)), \quad (2)$$

where the proportionality constant ensures that the distribution is normalized over the range of ages under consideration, for example, $t \in [6, 18]$ for school grades, or $t \in [0, 120]$ for durations representing the entire lifetime. In practice, although a truncated normal ensures the distribution is confined to an appropriate grade range, a normal without such restriction is a preferable model as it produces nearly identical results and is more mathematically tractable. The impact of substituting a normal distribution is explored in Section 3. Alternative unimodal distributions such as the Gamma family may also be used.

For a fixed vocabulary V of distinct words the acquisition age distributions for all words $w \in V$ are defined using $2|V|$ parameters [in the case of (2)]

$$\{(\mu_w, \sigma_w) : w \in V\}. \quad (3)$$

The parameters (3), which are the main objects of interest, can in principle be estimated from data using standard statistical techniques. Unfortunately, data containing explicit acquisition ages is very difficult to obtain reliably. Traditionally, estimating word acquisition was done based on interviewing adults regarding the age at which a word was acquired during childhood. Such data, however, may not be completely reliable due to the delay between the word acquisition ages and the interviews. Another source of difficulty is obtaining such data for a large representative group of people. Conducting such a survey takes substantial time and money. It is also hard to repeat such studies periodically which is necessary in order to get a contemporary model (as we describe later in the paper the age at which a word is acquired may evolve over time as society changes its emphases and values).

On the other hand, using modern web technology it is possible to reliably collect large quantities of documents paired with the ages of intended audiences. More specifically, such data may be automatically obtained by crawling textual resources intended for classroom use. In this paper, we demonstrate how to use such data to estimate the word acquisition parameters (3) and how to use the estimates to predict the readability ages of new unseen documents.

Traditionally, document readability has been defined in terms of the school grade level at which a large portion of the words have been acquired by most children (Chall and Dale 1995). Unfortunately, this definition is qualitative rather than quantitative and is not suitable for direct use in statistical modeling. We propose the following interpretation of that definition, which is made appropriate for quantitative studies by taking into account the inherent randomness in the acquisition process.

Definition 1. A document $d = (w_1, \ldots, w_m)$ is said to have (r, s)-readability level t if by age t no less than s percent of the words in d have been acquired each by no less than r percent of the population.

We denote by q_w the quantile function of the cdf corresponding to the acquisition distribution p_w. In other words, $q_w(r)$ represents the age at which r percent of the population $\mathcal{T}$ have acquired word w. Despite the fact that it does not have a closed form, it is a continuous and smooth function of the parameters μ_w, σ_w in (2) (assuming $\mathcal{T}$ is infinite) and can be tabulated before inference begins.

Following Definition 1 we define a readability model similar in form to the logistic Rasch model:

$$\log \frac{P(d \text{ is } (r,s)\text{-readable at age } t)}{1 - P(d \text{ is } (r,s)\text{-readable at age } t)} = \theta(q_d(r,s) - t) \quad (4)$$

or equivalently

$$P(d \text{ is } (r,s)\text{-readable at age } t) = \frac{\exp(\theta(q_d(r,s) - t))}{1 + \exp(\theta(q_d(r,s) - t))}, \quad (5)$$

where $q_d(r, s)$ is the s quantile of $\{q_{w_i}(r) : i = 1, \ldots, m\}$. However unlike the Rasch model, t does not have to be estimated as it takes on a prespecified range of grade levels or ages.

Figure 1 illustrates some of the concepts discussed above. It describes applying (r, s)-readability to a document consisting of 5 words with $r = 0.8$ and $s = 0.7$. The probability density functions p_w for the five words appear as dashed lines. Note that four words are typically acquired in grades 3–5 while one word is typically acquired at the later grades of 9–11. We illustrate the function $q_d(r, s)$ by plotting its cdf using as a solid piecewise constant function. The horizontal line indicates that grade 5.9 corresponds to the 0.7-quantile of $\{q_{w_i}(0.8) : i = 1, \ldots, m\}$.

In other words, the odds ratio of a document d being (r, s)-readable increases exponentially with $q_d(r, s)$ at a rate determined by the parameter θ. The parameter r determines what it means for a word to be acquired and is typically considered to be a high value such as 0.8. The parameter s determines how many of the document words need to be acquired for it to be readable. It can be set to a high value such as 0.9 if a very precise understanding is required for readability but can be substantially reduced when a more modest definition of readability applies.

In practice, predicting the readability of a novel document relies on an estimate of the sth quantile of the distribution of

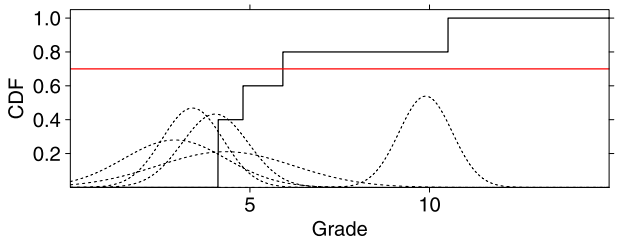

Figure 1. The figure describes applying (r, s)-readability to a document consisting of 5 words with $r = 0.8$ and $s = 0.7$. The probability density functions p_w for the five words appear as dashed lines. Note that four words are normally acquired in grades 3–5 while one word is typically acquired at the later grades of 9–11. We illustrate the function $q_d(r, s)$ by plotting its cdf using as a solid piecewise constant function. The horizontal line indicates that grade 5.9 corresponds to the 70th-quantile of $\{q_{w_i}(0.8) : i = 1, \ldots, m\}$. The online version of this figure is in color.

acquisition ages within the document which is performed using the sth order statistic. Assuming the normal distribution, we have that a word is acquired by r percent of the population at age

$$q_{w_i}(r) = \mu_{w_i} + \Phi^{-1}(r)\sigma_{w_i}. \quad (6)$$

To investigate the distribution of the predicted grade level of a document, we assume that the acquisition age parameters μ_w, σ_w corresponding to different words are drawn from Gamma distributions $\mu_w \sim G(\alpha_1, \beta_1)$ and $\sigma_w \sim G(\alpha_2, \beta_2)$. Noting that $\Phi^{-1}(r)\sigma_w \sim G(\alpha_2, \beta_2^*)$ where $\beta_2^* = \Phi^{-1}(r)\beta_2$, we can write the distribution of the acquisition ages, q_w, within a single document as the following convolution

$$f_Q(q) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta_1^{\alpha_1}\beta_2^{*\alpha_2}} q^{\alpha_1+\alpha_2-1} e^{-q/\beta_2^*} \times \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1} e^{((\beta_1-\beta_2^*)tq)/(\beta_1\beta_2^*)}\, dt \quad (7)$$

which reverts to the Gamma distribution when $\beta_1 = \beta_2^*$.

The distribution of the sth quantile of (7), which amounts to (r, s)-readability of documents, can be analyzed by combining f_Q above with a standard formula by David and Nagaraja (2003) for normal approximation of order statistics

$$q_d(r, s) \approx \mathrm{N}\left(F_Q^{-1}(s), \frac{s(1-s)}{m[f_Q(F_Q^{-1}(s))]^2}\right), \quad (8)$$

where m is the document length and F_Q is the cdf corresponding to (7). In practice this result allows the uncertainty in predicted readability level to be expressed as a function of document length, that is, we can be more certain about the grade level assigned to a longer document.

An example of the relationship between document length and confidence interval width of readability prediction is shown in Figure 2. For this illustration the distribution of acquisition ages corresponds to those in a randomly selected 1577 word document written for a 7th grade audience taken from the Web 1–12 corpus. The parameters for the acquisition age distributions were inferred using the method described in Section 3. Confidence intervals were then constructed in two different ways: (1) Gamma distributions for μ_w and σ_w were fit via maximum likelihood as described in the preceding paragraph, and (2) the empirical distribution of the 1577 inferred acquisition ages was used to directly determine the distribution f_Q. Finally, the 95% confidence intervals for the sth quantile of a document of a fixed length were created by generating 1000 Monte Carlo samples of size equal to document length. For example, for a document length of 150 words 1000 samples of size 150 were generated from f_Q, then the sth percentile for each was calculated. Lastly the distribution of the 1000 estimates was used to construct 95% confidence intervals.

Figure 2 contrasts the CI widths for model based intervals and empirical intervals. Interestingly, in both cases documents containing more than 100 words provide CI widths shorter than 1 year. This finding is remarkable since it provides empirical support for the long-standing "rule of thumb" in computational linguistics that readability measures become unreliable for passages of less than 100 words (Fry 1990).

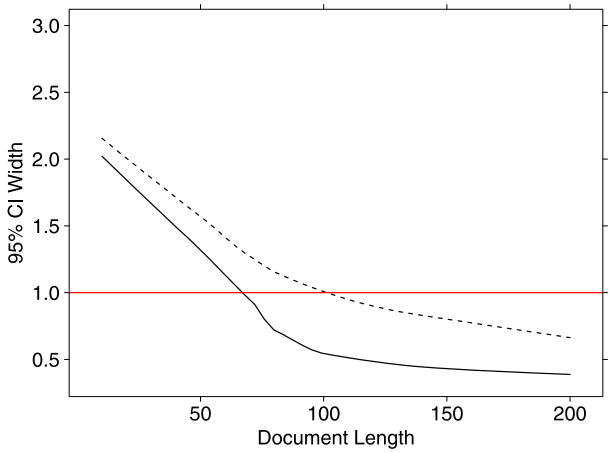

Figure 2. A comparison of model (dashed) versus empirical (solid) 95% confidence interval widths as a function of document length. A readibility definition of $r = 0.9$ and $s = 0.7$ was used based on typical inferred r, s values across our 3 corpora. The x-axis corresponds to document length, while the y-axis is the width of the 95% CI interval. The online version of this figure is in color.

3. EXPERIMENTAL RESULTS

In our experiments we used three readability datasets. The corpora were compiled by crawling web pages containing documents authored for audiences of specific grade levels. The first corpus contains 374 documents, with each document written for a particular school grade level in the range 1–12. The second corpus contains 1780 documents with each document authored for a particular grade in the range 2–5. The third is a collection of 215 documents with grade levels ranging from 1–6. The grade levels in all corpora correspond to United States school grades explicitly stated by the author or the classroom level where the documents were used. The pages were drawn from a wide range of subject areas, including history, science, geography, and fictional short stories. Additional details concerning the data and its collection may be found in the Appendix.

In the three sets of experiments that follow we used maximum likelihood to estimate the model parameters $\{(\mu_w, \sigma_w^2) : w \in V\}$, r, s, and θ for the readability Model (5) with a normal distribution for p_w in (2). Each document in the readability corpora is labeled with a specific grade level, which may be considered to represent the earliest grade at which that document is readable. This can be translated into a vector of length equal to the number of grade levels, for example, the readability of a fifth-grade text, $t_i = 5$ is described by $(0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1)$ indicating it is not readable in grades 1–4 and is readable at grades 5–12. Based on this representation, we estimate the parameters by maximizing the likelihood function over all documents in a corpus of size N.

$$L = \prod_{i=1}^{N}\prod_{t=1}^{12}(1 - P(d_i \text{ is readable at grade } t))^{I(t<t_i)} \times P(d_i \text{ is readable at grade } t)^{1-I(t\geq t_i)}, \quad (9)$$

where t_i indicates the grade level of document i, ranging from 1 to 12. The probability function above is the one defined in (4) using the normal acquisition distribution.

We note that due to the discreteness of the set $\{q_{w_i}(r) : i = 1, \ldots, m\}$, neither $q_d(r, s)$ nor the loglikelihood of (4) are differentiable in the parameters (3). This raises some practical difficulties with respect to the computational maximization of the likelihood and subsequent estimation of (3). However, for long documents containing a large number of words, $q_d(r, s)$ is approximately smooth which motivates a maximum likelihood procedure using gradient descent on a smoothed version of $q_d(r, s)$. Alternative optimization techniques which do not require smoothness may also be used. More information on iterative methods for nonlinear optimization of smooth and nonsmooth functions may be found in (Gill, Murray, and Wright 1981 and Boyd and Vandenberghe 2004).

We report below three experimental studies. The first examines the word acquisition distributions that were estimated based on the three corpora. The second examines predictions made by the model with respect to readability of new unseen documents. The third contains an experiment in predicting a local readability measure within a long heterogeneous document.

3.1 Estimation of the Word Acquisition Distributions

As mentioned above, we analyzed a corpus of documents tagged with the grade levels of the intended audience. Figure 3 displays the relatively uniform distribution over readability levels (top left) and the rather skewed distribution over document lengths (top right). Some documents had less than 100 words while others had more than 2500 words. The bottom left panel displays a box-plot of document lengths. As expected, document lengths vary substantially among the different grade levels with documents written for lower grade levels often being much shorter than those written for higher grades.

An empirical perspective on AoA distributions can be gained by examining the grade level at which words first appear in the corpus. Figure 3 (bottom right) contains a histogram of these grade levels. The shape of the histogram indicates a relatively uniform vocabulary growth, with the exception of slower vocabulary growth for grades 1–2.

Figure 4 displays the inferred acquisition age distributions and empirical word appearances of three words: `thought` (left), `multitude` (middle), and `assimilation` (right). In these plots, the empirical cdf of word appearances is indicated

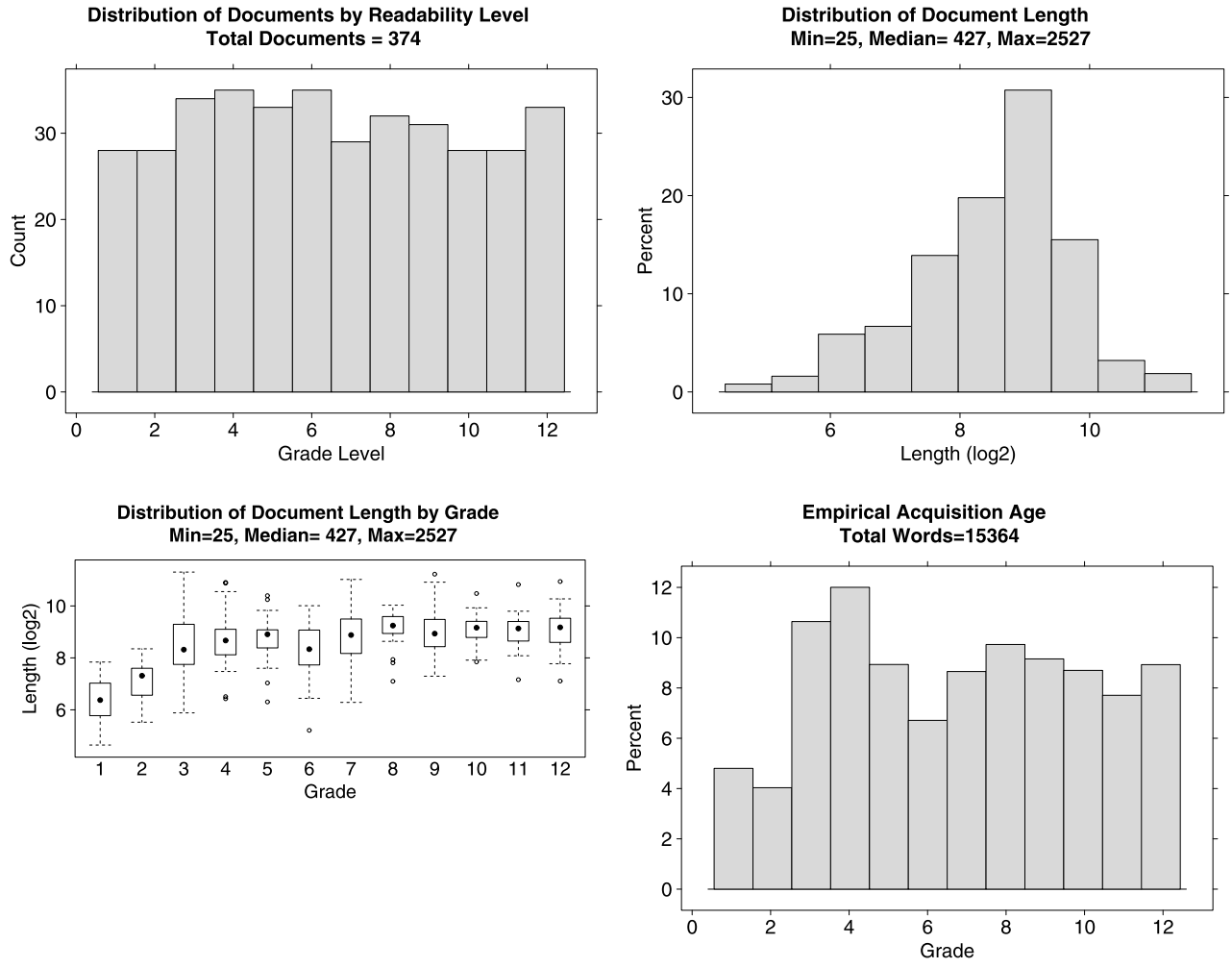

Figure 3. Top row: The corpus has a relatively uniform distribution over grade levels (left). The distribution of document length varies widely, ranging from less than 100 to more than 2500 (right). Bottom row: Document lengths vary substantially among the different grade levels. Documents written for lower grades are often much shorter than those written for higher grades (left). The histogram of grades in which words first appeared is nonuniform (right).

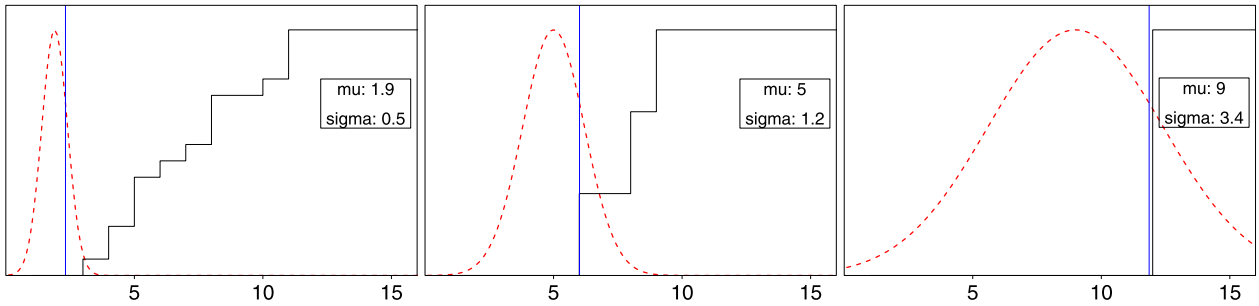

Figure 4. A comparison of empirical word appearances and AoA distributions for three words: `thought` (left), `multitude` (middle), and `assimilation` (right). The empirical cdf of word appearances appears as a piecewise constant line and the estimated pdf of the AoA is indicated by the dashed curve with its 0.8 quantile indicated by a vertical line. As expected, the distribution of the AoA tends to place more mass on lower grade levels relative to the empirical distribution of word appearances. The online version of this figure is in color.

by a piecewise constant line while the probability density function of the estimated AoA distribution is indicated by a dashed line. The empirical cdf of appearances for an individual word is determined by tabulating the total number of times the word appears in each grade across all documents in the corpus. The vertical line indicates the 0.8 quantile of the AoA distribution which corresponds to the grade by which 80% of the children have acquired the word.

The word `assimilation` appears in two documents having 12th grade readability. The high grade level of these documents results in a high estimated acquisition age and the paucity of observations leads to a large uncertainty in this estimate as seen by the variance of the acquisition age distribution. The word `thought` appears several times in multiple grades. It is first observed in the 3rd and 4th grades resulting in an estimated acquisition age falling slightly below the third grade. The variance of this acquisition distribution is relatively small due to the frequent use of this word. The empirical cdf shows that `multitude` is used in grades 6, 8, and 9. Relative to `thought` and `assimilation` the word `multitude` was used less and more frequently respectively, which leads to an acquisition age distribution with a larger variance than that of `thought` and smaller than that of `assimilation`.

The relationship in Figure 4 between the empirical word appearances and the acquisition age distribution demonstrates the following behavior: (a) The variance of the acquisition age distribution goes down as the word appears in more documents, and (b) the mean of the AoA distribution tends to be lower than the mean of the empirical word appearance distribution, and in many cases even smaller than the first grade in which the word appeared. This is to be expected since authors use specific words only after they believe they were acquired by a large portion of the intended audience.

In order to assess the impact of parametrizing the age of acquisition distribution with a normal instead of a truncated normal, we performed inference on the same set of words substituting the latter. Figure 5 shows a plot of μ_w where the x-axis represents the value calculated under the normal distribution and the y-axis the truncated normal. At the lowest grade levels the truncated normal distribution tends to produce slightly higher estimated mean acquisition ages, while this impact diminishes as the age of acquisition increases. These trends can be attributed to the small amount of probability assigned to grade levels of less than zero under the normal distribution.

The quartiles of the acquisition age distributions for 10 words appear in Table 1. Words such as `thought`, `muffin`, `memory` have a median acquisition grade level of 1–2 while more complicated words such as `wrought`, `innovation`, `assimilate` have median acquisition grades of 6 of higher.

Although this work focuses on estimating a written age of acquisition, it is interesting to compare these results to related studies in the linguistic community concerning the oral acquisitions of words. These studies estimate the age at which a word is acquired for oral use based on interview processes with participating adults. We focus specifically on the seminal study of acquisition ages performed by Gilhooly and Logie (GL) (1980) and made available through the MRC psycholinguistic database (Coltheart 1981).

There are some substantial differences between these previous studies and our approach. We analyze the age of acquisition through document readability which leads to a written, rather than oral, notion of word acquisition. Furthermore, our

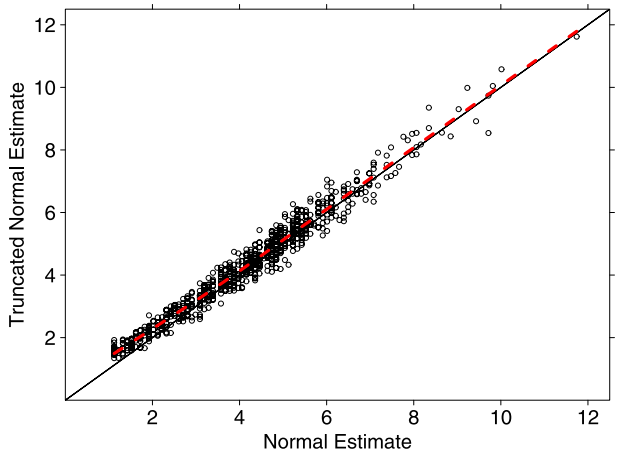

Figure 5. Comparing the inferred values for μ_w using the truncated normal (y-axis) and normal distribution (x-axis) shows this substitution has a relatively small impact on $(\mu_w, \sigma_w) : w \in V$. Fitting a lowess curve (dashed red) shows that the inferred truncated mean does tend to be slightly larger at low grade levels, while on average the difference is approximately 0.15 grade levels and has a MAE of 0.28 grade levels. The online version of this figure is in color.

Table 1. Although acquisition age can be represented by a single number such as the 0.8 quantile, a broader perspective can be gained by examining acquisition age by its distribution. The quartiles provide some insight into the nature of these distributions

Word	1st quartile	median	3rd quartile
muffin	1.0	1.3	1.6
thought	1.6	1.9	2.2
memory	1.8	2.2	2.6
transform	2.6	3.1	3.6
perfectly	3.9	4.8	5.7
multitude	4.2	5.0	5.8
artisan	4.4	5.2	6.0
wrought	5.2	6.1	7.0
innovation	5.5	6.5	7.5
assimilate	6.7	9.0	11.3

estimates are based on documents written with a specific audience in mind, while the previous studies are based on interviewing adults regarding their childhood word acquisition process which may be less reliable due to the age difference between the acquisition and the interview. Finally, the GL study was performed in the late 1970s while our study uses more contemporary internet data. It is reasonable to expect that the grade level at which some words are acquired to have changed over the past three decades.

Although substantial differences exist between this study and the oral studies some correlation is present between the written ages and both the GL ($r^2 = 0.34$) and more recent AoA list of Cortese and Khanna (Cortese and Khanna 2008) ($r^2 = 0.43$). As illustrated in Figure 6 the acquisition ages obtained from written readability data tend to be higher than the GL study. This is to be expected as oral acquisition ages (which the GL study focuses on) tend to be lower than written acquisition ages. As indicated by Figure 6 this tendency diminishes as the grade level increases which may represent an increase in the likelihood that a word is learned in the written form at the same time it is acquired orally.

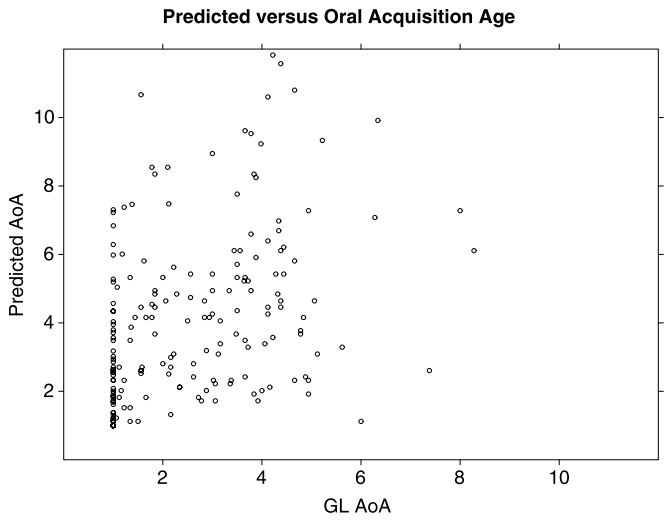

Figure 6. A scatter plot ($s = 0.8$, $\theta = 50$, $n = 200$) of predicted age of acquisition versus Gilhooly and Logie's values reveals the tendency for the written estimate to exceed the oral estimate ($r^2 = 0.34$).

In some interesting cases, the ages obtained from readability data are actually lower than the ages reported in the older oral studies. Two plausible explanations for this are a shift in educational standards or a change in social standards. Approximately 30 years have passed since Gilhooly and Logie's study was conducted. During these three decades, social and educational standards have changed. For example, grade school curricula have changed to incorporate the fundamentals of the scientific method resulting in appearances of words such as hypothesis, conclusion, engineer earlier than the oral GL acquisition ages. Specifically, the acquisition grade levels for these words have decreased by 4.2, 2.4, and 0.6 respectively. Similarly a change in societal standards emphasizing drug awareness may be observed in the word drug which appeared in writing 0.94 grades earlier than the oral AoA of the GL study. The newer Bristol Norm study confirms this observation as it predicts a decrease in grade level of 0.88 over GL as well.

3.2 Global Readability Prediction

We have seen in the previous section how the acquisition ages are estimated from document readability data. Once the acquisition ages are available, whether estimated statistically from data or obtained from a survey such as the GL survey, they may be used to predict the grade level of novel documents. Such predictions are useful in information system engineering where a common task is to return a list of relevant documents in response to a query. Matching the readability level in addition to topic relevance may help to achieve a better search experience for children or nonnative speakers of English (or any other language). The predictions may also be used to verify whether materials assigned for classroom studies are appropriate for use at different grade levels.

Specifically, our *fixed threshold model* is based directly on Equation (5) and predicts readability level t^* for a novel document d if it is the minimal grade for which readability is established:

$$t^* = \min\{t : P(d \text{ is readable at age } t) \geq \beta\}, \tag{10}$$

where β is some parameter describing the strictness of the readability requirement.

In our experience, using a constant β as above leads to a biased predictor, for example, $\beta = 0.5$ leads to accurate predictions at lower grades and under-prediction at high grades. This effect can be attributed to handling of words that have never been seen in the training text, that is, high grade level words appear less frequently than low grade level words and therefore have a disproportionate impact on predicted grade level. Figure 7 clearly shows the negative bias which increases linearly under the fixed threshold rule (left) and the correction provided by a dynamic threshold rule (right).

Therefore, we employ a *dynamic threshold model* based on (10) with β being a function of t rather than a fixed num-

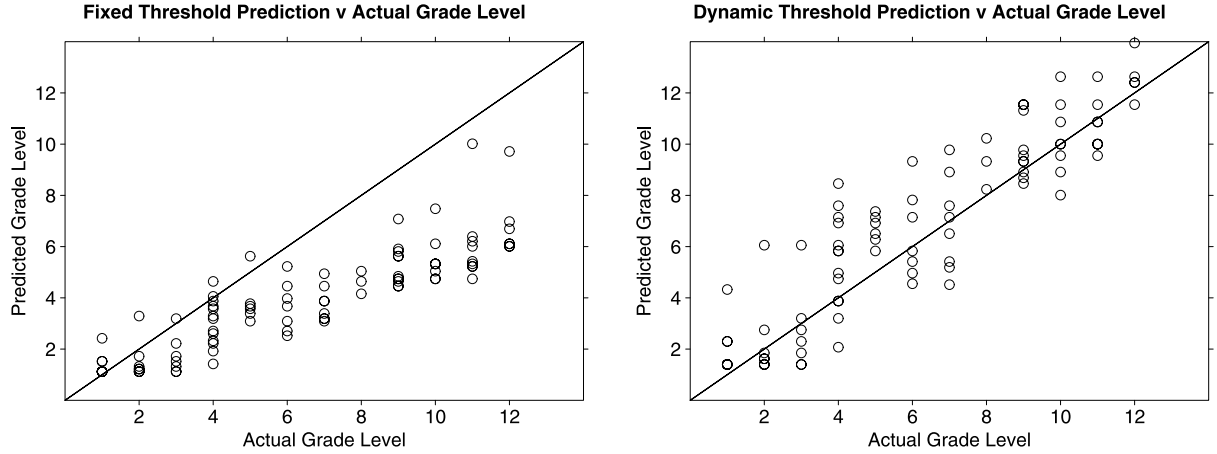

Figure 7. The scatter plots demonstrates the strong relationship between predicted and actual global readability levels using the fixed (left) and dynamic (right) prediction thresholds.

ber. Specifically, we used

$$t^* = \min\{t : P(d \text{ is readable at age } t) \geq \beta(t)\} \tag{11}$$

$$= \min\left\{t : \frac{\exp(\theta(q_d(r,s)-t))}{1+\exp(\theta(q_d(r,s)-t))} \geq \frac{\exp(at-b)}{1+\exp(at-b)}\right\} \tag{12}$$

$$= \frac{\theta q_d(r,s) - b}{a+\theta}, \tag{13}$$

where $\beta(t) = \frac{\exp(at-b)}{1+\exp(at-b)}$. The range of the resulting $\beta(t)$ is typically 0.5 in lower grades, increasing to 0.9 in higher grades. An alternative interpretation of the prediction is to observe that the objective is to predict the r, s-readability of the document which may be estimated by $\hat{q}_d(r,s)$. As shown in the figure this estimator has a negative bias, therefore fitting a linear model to the bias provides the correction, $a'\hat{q}_d(r,s) + b'$.

Implementation of this predictive model requires estimating word specific parameters $\{(\mu_w, \sigma_w^2) : w \in V\}$, the model parameter θ, and dynamic threshold parameters a, b. In order to perform this prediction it is necessary to split the corpus into three components (A, B, C) as two training steps and a testing step are required: fitting the readability model (A), fitting the bias model (B), and evaluating performance (C). In the first step word specific parameters are estimated as described at the beginning of the section. Secondly, note the model parameter θ is not identifiable in the predictive setting as a result of the three-staged training-testing paradigm. Refering to Equation (11) it is clear that estimating a, b after θ will make the initial estimate of θ irrelevent. Finally, the dynamic threshold parameters, a, b, can be estimated by fitting a linear model via maximum likelihood to the residuals based on the fixed threshold model of the first training step.

Each of the following experiments focused on measuring an algorithm's capacity to predict readability using the metric of mean absolute error (MAE). The estimation of the dynamic threshold prediction model required splitting the data into the three subsets A–B–C according to the proportions 75–15–10, that is, a 90% training and 10% test split. The three-component procedure was also used for evaluating the support vector regression (SVR) based on AoA percentiles as some of the data had to be used to infer these percentiles, while every method utilizing the standard training-test procedure was evaluated over a 90–10 split. This training-test procedure ensured that each algorithm had the same overall proportion of the data allocated to training regardless of the number of steps involved. All confidence intervals on performance estimates were then obtained by randomly sampling 100 assignments of documents into the training–training–test or training–test partitions.

A first step in understanding the dynamic threshold prediction rule is to examine its performance under a range of parameter values. Figure 8 (left) shows the correlation and MAE of the predictions with the corresponding 90% confidence intervals as functions of the quantile parameter s. The best value of s for both quantities is around 0.6 (0.65 for the MAE). The best correlation is 0.89 which is notably close to 1. We compared the predictions of the dynamic threshold model (11) to two standard classifiers: naive Bayes and SVR. SVR was applied twice using different sets of features—once with the document word

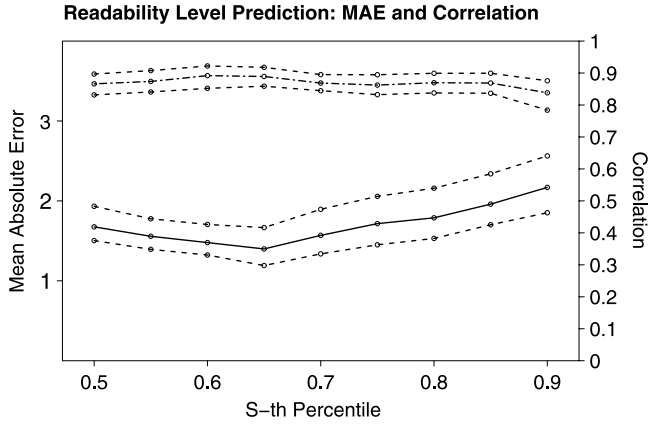

Figure 8. Mean absolute error (MAE) and correlation coefficient as functions of the quantile parameter s. The MAE is displayed as the solid line and is aligned with the left axis while the correlation is displayed as a dashed line and is aligned with the right axis. 90% bootstrap confidence intervals are displayed.

Table 2. A comparison of mean absolute error (MAE) across prediction algorithms shows the age of acquisition model compares favorably. The confidence bounds (LB,UB) were computed by repeating each model building procedure 100 times. Note that the Dynamic Threshold Model (GL) indicates the performance using AoA inferred from the GL survey

Prediction rule	MAE	LB	UB
Fixed threshold model	1.69	1.47	1.94
Dynamic threshold model	1.40	1.19	1.67
Dynamic threshold model (GL)	1.41	1.27	1.62
Naive Bayes	1.94	1.67	2.19
SVR (word frequency)	1.83	1.66	2.03
SVR (AoA percentiles)	1.36	1.22	1.58
Grade 6	2.92	–	–

frequencies as features and once with the estimated AoA percentiles for the document words as features. The MAE for the 4 predictors and their 90% confidence intervals are displayed in Table 2. Prediction methods based on word frequencies performed substantially worse than using the estimated acquisition distributions; specifically the naive Bayes model obtained a relatively poor 1.94 MAE, while the SVR and dynamic threshold model produced a MAE near 1.4. These all compare favorably to the 2.92 MAE obtained by using a constant prediction rule of always predicting grade 6.

High quality readability prediction is a worthwhile result in itself; however, we can also use the prediction mechanism to study the validity of Definition 1 and the associated probability model. We do so by applying other predictive algorithms using the inferred acquisition age distribution for each document as the predictor variables and comparing the MAE with the MAE obtained by the estimated dynamic threshold model. In particular, we examine the performance of support vector regression (SVR) using the estimated AoA percentiles for each document as predictor variables. The results displayed in Table 2 show that SVR and the dynamic threshold model perform similarly well allowing us to conclude that Definition 1 forms the basis for a suitable model for readability prediction.

It is also interesting to consider predicting readability using acquisition ages obtained in surveys rather than the ages obtained from the maximum likelihood estimation. Specifically, we use the GL age of acquisition norms which are completely independent of the corpus. The intersection of AoA norm data and the corpus is 1217 words; additionally, the highest grade level associated with the norms is the eighth. Therefore, we did restrict the experiments to documents of readability grade eight or less. A difference in prediction using normed acquisition ages is that normed acquisition age is typically described by a single value as opposed to a distribution. When applying the prediction rule using AoA norms r is implicitly selected in the norming process as the result is a single value instead of a distribution. The dynamic threshold procedure was applied with the sth percentiles ranging from 92 to 100, a smaller range than that used for the inferred AoA although still with the same functional relationship. This difference is not surprising given the tendency for the inferred age of acquisition to exceed the normed value, the implicit r, and the limited intersection. The results displayed in Table 2 show that applying the dynamic threshold model with the GL survey based AoA data provides a MAE $= 1.41$ which is comparable to the performance using the inferred AoA. The degree of similarity in performance is somewhat surprising considering the number of differences in the AoA estimation, however the high quality of predictions based on survey data lends strong support to the validity of the assumptions made in Definition 1.

3.3 Local Readability Prediction

Longer documents such as books or movies may display a variety of local readability levels. This disparity can come from a number of sources, including different topics, chapters, or dialogue. Several previous readability studies have recognized this behavior including some of the earliest readability scores such as Flesch (1948), Dale–Chall (1948), and Fry (1968). These methods typically sampled passages throughout a text, and then combined the readability levels of the passages to produce an overall readability score. We are interested in a local readability estimate due to its importance in document summarization, browsing, and general document understanding.

In order to extend our model to predict local readability we have applied a locally weighted version of our model. For simplicity we used a sliding window procedure equivalent to a local likelihood procedure with a constant kernel function. The width of the sliding window or kernel, which we denote by h, corresponds to the number of words that are considered as around the estimation locus. As the window slides from the beginning of the document to its end, the dynamic threshold model (11) is applied resulting in a sequence of local readability estimates. The width of the window determines the degree of locality and can be viewed as a parameter controlling the degree of smoothing. A narrow window emphasizes local behavior such as a specific conversation, while a broad one captures more long term trends.

We investigated a popular action film, "The Matrix: Reloaded," to gain perspective on the effectiveness of local readability prediction. This movie was chosen because it displayed a wide variety of dialogue ranging from simple during action scenes to highly complex in abstract futuristic scenes. In a series of experiments we explored a range of window widths. We found that $h = 100$, as used in Figure 9, was narrow enough to capture the conversation level detail, yet wide enough to avoid wild fluctuations associated with extremely local behavior. The determination of optimal window width must be a heuristic procedure at a certain level, as training data for local readability do not exist; however, the combination of knowledge of the "Matrix," our result relating confidence interval width and document length in Section 2, and the well-known rule of thumb regarding passages of length 100 combine to produce a seemingly reasonable representation of local readability. Figure 9 illustrates that local readability does capture the variability in readability across the various scenes. In particular, the low area A corresponds to action scenes and simple dialogues. Peaks B and C correspond to the complex dialogues with the Merovingian and with the architect, respectively.

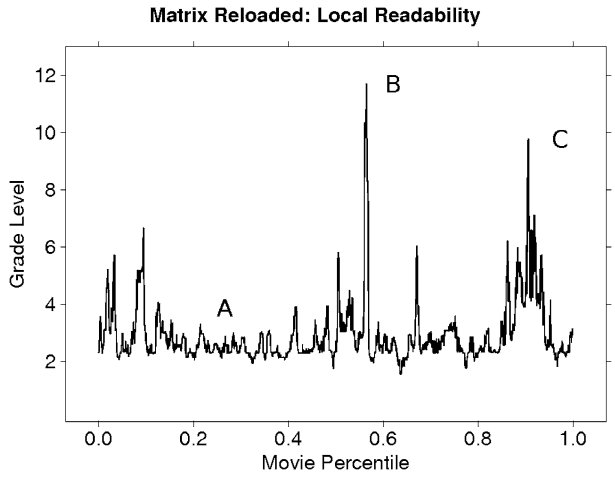

Figure 9. The local readability of "The Matrix Reloaded" is displayed using a moving window approach. The complexity of the dialogue shifts dramatically in the course of the movie. The low area *A* corresponds to action scenes and simple dialogues. Peaks *B* and *C* correspond to the complex dialogues with the Merovingian and with the architect, respectively.

4. RELATED WORK

Age of acquisition for word reading and understanding has been extensively studied as a learning factor in the psycholinguistics literature, where AoA norms have been obtained using surveys. Examples of relevant literature are Gilhooly and Logie (1980) and Zevin and Seidenberg (2002). Our approach differs by connecting AoA to readability through Definition 1 and using readability data to estimate AoA norms from large amounts of authentic language data. A recent related study is that by Crossley, McCarthy, and McNamara (2007) who used AoA as a factor in analyzing reading difficulty to help discriminate between authentic and simplified texts for second-language readers.

Over the past 15 years, there has been renewed interest in corpus-based statistical models for readability prediction. One example is the popular Lexile measure (Stenner 1996) which uses word frequency statistics from a large English corpus. Collins-Thompson and Callan (2005) introduced a new approach based on statistical language modeling, treating a document as a mixture of language models for individual grades. Further recent refinements in methods for readability prediction include using machine learning methods such as Support Vector Machines (Schwarm and Ostendorf 2005), log-linear models (Heilman, Collins-Thompson, and Eskenazi 2008), *k*-NN classifiers, and combining semantic and grammatical features (Heilman et al. 2007). There has been very little work on considering readability as a local quantity to be predicted and most studies aggregate document information to construct a single global measure.

5. DISCUSSION

While there have been several recent studies regarding word acquisition and readability our work is the first to provide a quantitative connection between these two concepts in a statistically meaningful way. The core assumption that we make is Definition 1 which is consistent with standard readability definitions, for example, Chall and Dale (1995) and states that document readability level is determined by most people understanding most words. Although this definition does not capture the full complexity associated with readability, the persistence of this relationship in the literature, the fact that previous work attributes only 10% of the variability in readability level to a syntactic component (Chall and Dale 1995), and the correlation between the inferred ages of acquisition and survey based measures provide a degree of reassurance with regard to this core assumption. Experimentally, readability grade level predictions using the GL norms performed very similarly to the inferred norms which also lends support to the assumption of Definition 1. The precise form of Definition 1 is not unlike the probably approximately correct (PAC) framework of Valiant (1984) which states that a concept is learned if it is achieved approximately with high probability. As such, it is suitable for complex probabilistic relationships and for statistical learning from data.

The connection between word acquisition and readability is both intuitive and useful. It allows two degrees of freedom s and r to handle situations where different readability notions exist. Experiments validate the model and demonstrate interesting trends in word acquisitions as compared to older oral acquisition studies. From an engineering standpoint, it allows estimation of word acquisition parameters directly from web data as opposed to user studies and interviews.

Experimental results show that the proposed model is also effective in terms of predicting readability level of documents. It compares favorably to naive Bayes and support vector regression, the latter being one of the strongest regression baselines. Of particular note are our observations regarding the necessity of a dynamic threshold with $\beta(t)$ being a monotonic increasing function, and the extension of global readability prediction to local readability prediction in Section 3.3. The proposed framework offers the flexibility to serve as an ensemble estimator allowing a single document to contribute multiple observations via readability scores computed across existing methodologies.

APPENDIX: DATASET DESCRIPTION AND EXPERIMENTAL SETUP

Corpus 1 was derived from a set of 374 web pages gathered in a three-week period during the years 2001–2002. Each page had been assigned a level in $\{1, \ldots, 12\}$ corresponding to U.S. school grades based on the explicit grade stated by the author or the classroom level where the document was acquired. The pages were drawn from a wide range of subject areas, including history, science, geography, and fictional short stories. The original documents were a mixture of html, pdf, and text files and were converted to lowercase plain text using standard parsers. Overall, there were 15,833 unique words (types) across all documents and a total of 183,389 words (tokens).

Corpus 2, the weekly reader dataset, was obtained by crawling the Weekly Reader commercial website after receiving special permission. The readability dataset contains a total of 1780 documents, with four readability levels ranging from 2 to 5 indicating the school grade levels of the intended audience. A total of 788 documents with readability between grades 2 and 5 and having length greater than 50 words were selected from 1780 documents.

Corpus 3, the Reading A–Z dataset, contains a set of 215 documents spanning grades 1 through 6. These documents were collected by crawling the corporate website Reading Arul-Z.com during the year 2003.

An important practical question is how to handle words that were not observed in any of the given documents. We use a heuristic smoothing procedure which works well in many language applications: we assumed that such words appeared 0.1 times in each grade and proceeded with the standard maximum likelihood estimate over the modified counts.

[Received May 2009. Revised July 2010.]

REFERENCES

Baeza-Yates, R., and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Boston: Addison Wesley. [21]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press. [24]

Chall, J. S., and Dale, E. (1995), *Readability Revisited: The New Dale–Chall Readability Formula*, Brookline, MA: Brookline Books. [22,29]

Collins-Thompson, K., and Callan, J. (2005), "Predicting Reading Difficulty With Statistical Language Models," *Journal of the American Society for Information Science and Technology*, 56, 598–605. [29]

Coltheart, M. (1981), "The MRC Psycholinguistic Database," *Quarterly Journal of Experimental Psychology*, 33A, 497–505. [25]

Cortese, M., and Khanna, M. (2008), "Age Acquisition Ratings for 3000 Monosyllabic Words," *Behavior Research Methods*, 40, 791–794. [26]

Crossley, S. A., McCarthy, P. M., and McNamara, D. S. (2007), "Discriminating Between Second Language Learning Text-Types," in *Proc. of the Twentieth International Florida Artificial Intelligence Research Society Conference*, Menlo Park, CA: AAAI Press. [29]

Dale, E., and Chall, J. (1948), "A Formula for Predicting Readability," *Educational Research Bulletin*, 27, 11–20. [28]

David, H. A., and Nagaraja, H. N. (2003), *Order Statistics*, Marblehead, MA: Wiley. [23]

Flesch, R. (1948), "A New Readability Yardstick," *Journal of Applied Psychology*, 32, 221–233. [28]

Fry, E. (1968), "A Readability Formula That Saves Time," *Journal of Reading*, 11, 513–516. [28]

——— (1990), "A Readability Formula for Short Passages," *Journal of Reading*, 33, 594–597. [23]

Gilhooly, K. J., and Logie, R. H. (1980), "Age of Acquisition, Imagery, Concreteness, Familiarity and Ambiguity Measures for 1944 Words," *Behaviour Research Methods and Instrumentation*, 12, 395–427. [25,29]

Gill, P. E., Murray, W., and Wright, M. H. (1981), *Practical Optimization*, Academic Press. [24]

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007), "Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts," in *Proceedings of the Human Language Technology Conference*, Columbus, OH. [29]

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008), "An Analysis of Statistical Models and Features for Reading Difficulty Prediction," in *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications*. [29]

Schwarm, S. E., and Ostendorf, M. (2005), "Reading Level Assessment Using Support Vector Machines and Statistical Language Models," in *Proceedings of the Association of Computational Linguistics*. [29]

Stenner, A. J. (1996), *Measuring Reading Comprehension With the Lexile Framework*, Durham, NC: Metametrics, Inc. [29]

Valiant, L. G. (1984), "A Theory of the Learnable," *Communications of the ACM*, 27, 1134–1142. [29]

Zevin, J. D., and Seidenberg, M. S. (2002), "Age of Acquisition Effects in Word Reading and Other Tasks," *Journal of Memory and Language*, 47, 1–29. [29]