Statistical Estimation of Word Acquisition with Application to Readability Prediction

Paul Kidwell
Department of Statistics
Purdue University
West Lafayette, IN
kidwellpaul@gmail.com

Guy Lebanon
College of Computing
Georgia Institute of Technology
Atlanta, GA
lebanon@cc.gatech.edu

Kevyn Collins-Thompson
Microsoft Research
Redmond, WA
kevynct@microsoft.com

Abstract

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. We present a novel statistical approach that can infer the distribution of a word's likely acquisition age automatically from authentic texts collected from the Web. We then show that combining these acquisition age distributions for all words in a document provides an effective semantic component for predicting reading difficulty of new texts. We also compare our automatically inferred acquisition ages with norms from existing oral studies, revealing interesting historical trends as well as differences between oral and written word acquisition processes.

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. 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 authored specifically for fourth graders are said to have readability grade level four.

We develop and evaluate a novel statistical model that draws a connection between document grade level readability and age acquisition distributions. Based on previous work in the area, we define a model for document readability using a logistic Rasch model and the quantiles of the acquisition age distributions. We then proceed to infer the age acquisition distributions for different words from document readability data collected by crawling the web.

We examine the inferred acquisition distributions from two perspectives. First, we analyze and contrast them with previous studies on oral word acquisition, revealing interesting historical trends as well as differences between oral and written word acquisition processes. Second, the inferred acquisition distributions serve as parameters for the readability model, which enables us to predict the readability level of novel documents.

To our knowledge, this is the first published study of a method to 'reverse-engineer' individual word acquisition statistics from graded texts. By obtaining such a fine-grained model of how language evolves over time, we obtain a new, rich source of semantic features for a document. The increasing amounts of content available from the Web and other sources also means that these flexible models of authentic usage can be easily adapted for different tasks and populations. Our work serves to complement the growing body of research using statistics and machine learning for language learning tasks, and has applications including predicting reading difficulty for Web pages and other non-traditional documents, reader-specific example and question generation for lexical practice in intelligent tutoring systems, and analysis tools for language learning research.

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 at which word w was

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 900–909,
Singapore, 6-7 August 2009.

acquired by the population. Existing AoA norm studies almost universally summarize AoA ratings in terms of two parameters: mean and standard deviation, ignoring higher-level moments such as skew. For direct comparison with these studies we follow this convention and thus our goal is to estimate AoA for a word w in terms of mean μ_w and standard deviation σ_w parameters using the (truncated) normal distribution

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

where the proportionality constant ensures that the distribution is normalized over the range of ages under consideration e.g., $t \in [6, 18]$ for school grades. It is important to note that our model is not restricted by the assumption of (1) and can be readily extended to the Gamma family of distributions, if modeling asymmetric spread in the distribution is appropriate.

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

$$\{(\mu_w, \sigma_w) : w \in V\}. \quad (2)$$

These parameters, 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. Explicit word acquisition data is based on interviewing adults regarding their age acquisition process during childhood and so may be unreliable and difficult to obtain for a large representative group of people.

On the other hand, it is possible to reliably collect large quantities of readability data defined as pairs of documents and ages of intended audience. As we demonstrate later in the paper, such data may be automatically obtained by crawling specialized resources on the Web. We demonstrate how to use such data to estimate the word acquisition parameters (2) and to use the estimates to predict future readability ages.

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). 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 $(1-\epsilon_1, 1-\epsilon_2)$-readability level t if by age t no less than $1-\epsilon_1$ percent of the words in d have been acquired each by no less than $1-\epsilon_2$ 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 (1) (assuming $\mathcal{T}$ is infinite) and can be tabulated before inference begins.

Following Definition 1 we define a logistic Rasch readability model:

$$\log \frac{P(d \text{ is } (s,r)\text{-readable at age } t)}{1 - P(d \text{ is } (s,r)\text{-readable at age } t)} = \theta(q_d(s,r) - t) \quad (3)$$

where $q_d(s,r)$ is the s quantile of $\{q_{w_i}(r) : i = 1, \ldots, m\}$. An equivalent formulation to (3) that makes the probability model more explicit is

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

In other words, the probability of a document d being (s,r)-readable increases exponentially with $q_d(s,r)$ which is the age at which s percent of the words in d have been acquired each by r percent of the population.

The parameter $r = 1 - \epsilon_2$ 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 $s = 1 - \epsilon_1$ 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 reduced when a more modest definition of readability applies.

We note that due to the discreteness of the set $\{q_{w_i}(r) : i = 1, \ldots, m\}$, neither $q_d(s,r)$ nor the loglikelihood are differentiable in the parameters (2). This raises some practical difficulties with respect to the computational maximization of the likelihood and subsequent estimation of (2). However, for long documents containing a large number of words, $q_d(s,r)$ is approximately smooth which motivates a maximum likelihood procedure using gradient descent on a smoothed version of

$q_d(s)$. Alternative optimization techniques which do not require smoothness may also be used.

In the case of a normal distribution (1) we have that a word is acquired by r percent of the population at age $w = \mu + \Phi^{-1}(r)\sigma$, where Φ is the cumulative distribution function (cdf) of the normal distribution. To investigate the distribution of acquisition ages we assume that the μ, σ parameters corresponding to different words in a document are drawn from Gamma distributions $\mu \sim G(\alpha_1, \beta_1)$ and $\sigma \sim G(\alpha_2, \beta_2)$. The normal and Gamma distributions are chosen in part because they are flexible enough to model many situations and also admit good statistical estimation theory. Noting that $\Phi^{-1}(r)\sigma \sim G(\alpha_2, \Phi^{-1}(r)\beta_2)$, we can write the distribution of the acquisition ages as the following convolution

$$f_W(w) = \frac{w^{\alpha_1+\alpha_2-1}e^{-w/\beta_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta_1^{\alpha_1}\beta_2^{\alpha_2}} * \int_0^1 \frac{t^{\alpha_1-1}e^{\frac{(\beta_1-\beta_2)tw}{\beta_1\beta_2}}}{(1-t)^{1-\alpha_2}}dt$$

which reverts to a Gamma when $\beta_1 = \beta_2$.

The distribution of the s-percentile of f_W, which amounts to (r, s)-readability of documents, can be analyzed by combining f_W above with a standard normal approximation of order statistics (e.g., (David and Nagaraja, 2003))

$$X_{\lfloor mp \rfloor} \sim N\left(F_W^{-1}(p), \frac{p(1-p)}{m[f_W(F_W^{-1}(p))]^2}\right)$$

where m is the document length and F_W is the cdf corresponding to f_W.

Figure 1 shows the relationship between document length and confidence interval (CI) width in readability prediction. It contrasts the CI widths for model based intervals and empirical intervals. In both cases, documents of lengths larger than 100 words provide CI widths shorter than 1 year. This finding is also noteworthy as it provides empirical support for the long-standing 'rule-of-thumb' that readability measures become unreliable for passages of less than 100 words (Fry, 1990).

3 Experimental Results

Our experimental study is divided into three parts. The first part examines the word acquisition distributions that were estimated based on readability data. The second part compares the estimated

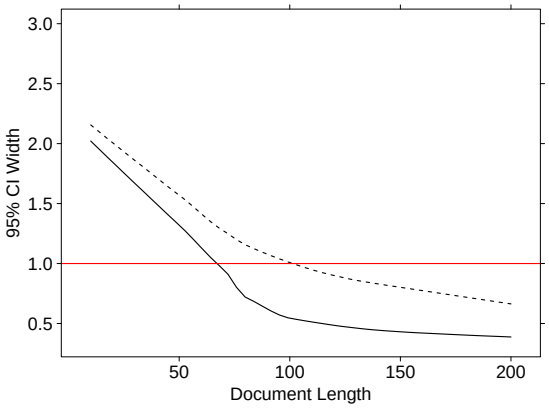

Figure 1: A comparison of model (dashed) vs. empirical (solid) 95% confidence interval widths as a function of document length ($r = 0.9$ and $s = 0.7$). CI widths were computed using 1000 Monte Carlo samples generated from the f_W model fit to the data and from the empirical distribution. Word distributions correspond to a 1577 word document written for a 7th grade audience taken from the Web 1-12 corpus.

(written) acquisition ages with oral acquisition ages obtained from interview studies reported in the literature. The third part focuses on using the estimated word acquisition distributions to predict document readability. These three experimental studies are described in the three subsections below.

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 Web 1-12 data contains 373 documents, with each document written for a particular school grade level in the range 1-12. The Weekly Reader (WR) dataset, was obtained by crawling the commercial website www.wrtoolkit.com after receiving special permission. That dataset contains a total of 1780 documents, with 4 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. The Reading A-Z dataset, contains a set of 215 documents was obtained from Reading A-Z.com, spanning grade 1 through grade 6.

The grade levels in these three corpora, which correspond to US school grades, were either explicitly specified by the organization or authors

who created the text, or implicit in the classroom curriculum page where the document was acquired. The pages were drawn from a wide range of subject areas, including history, science, geography, and fiction.

To reduce the possibility of overfitting, we used a common feature selection technique of eliminating words appearing in less than 4 documents. In the experiments we used maximum likelihood to estimate the model parameters $\{(\mu_w, \sigma_w^2) : w \in V\}$ for the Rasch model (3). The maximum likelihood was obtained using a non-smooth coordinate descent procedure.

3.1 Estimation of Word Acquisition Distributions

Figure 2 displays the inferred age acquisition distributions and empirical word appearances of three words: `thought` (left), `multitude` (middle), and `assimilate` (right). In these plots, the empirical cdf of word appearances is indicated by a piecewise constant line while the probability density function of the estimated AoA distribution is indicated by a dashed line. 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 2 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 1st grade and not again until the 4th grade resulting in an estimated acquisition age falling between the two. 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 2 between the empirical word appearances and the age acquisition distribution demonstrates the following behavior: (a) The variance of the age acquisition 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 as authors use specific words only after they believe the words were acquired by a large portion of the intended audience.

3.2 Comparison with Oral Studies

Among the related work in the linguistic community, are several studies concerning oral acquisitions of words. These studies estimate the age at which a word is acquired for oral use based an 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 database (Coltheart, 1981).

There are some substantial differences between these previous studies and our approach. We analyze the age acquisition process through document readability which leads to a written, rather than oral, notion of word acquisition. Furthermore, our 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 is arguably 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 contemporary internet data. Conceivably, the word acquisition process changed over the past 3 decades.

Despite these differences, it is interesting to contrast our inferred age acquisitions with the GL study and consider the differences and similarities. Figure 3 displays the relationship between the GL age of acquisition (AoA) and the acquisition ages obtained from readability data based on the $s = 0.8$ quantile. Some correlation is present ($r^2 = 0.34$) but the two measures differ considerably. As expected, the acquisition ages obtained from written readability data tend to be higher than the oral studies. The distributions of differences between the GL acquisition ages and the ones inferred from the readability data appears in Figure 4.

Comparing the acquisition ages obtained from readability data to the GL study results in a mean absolute error of 0.9 to 1.5, depending on the spe-

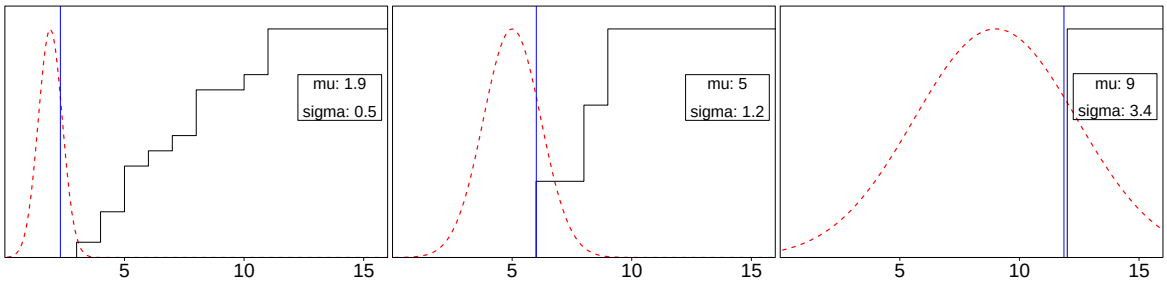

Figure 2: A comparison of empirical word appearances and AoA distributions for three words: `thought` (left), `multitude` (middle), and `assimilation` (right). The empirical cdf of word appearances appears as a piecewise constant line and the estimated pdf is indicated by the dashed curve with its 0.8 quantile indicated by a vertical line.

cific value of the Rasch parameter θ. Interestingly, the tendency for the written acquisition age to exceed the oral one diminishes as the grade level increases. This represents the notion that at higher grades words are acquired in both oral and written senses at the same age.

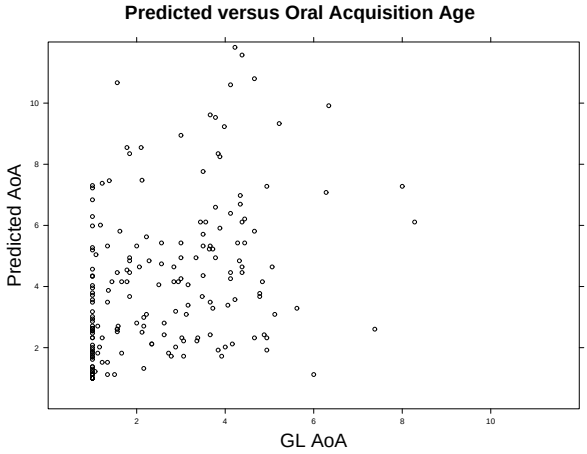

Figure 3: A scatter plot ($s = 80$, $n = 50$) of predicted age of acquisition versus Gilhooly and Logie's values reveals the tendency for the written estimate to exceed the oral estimate ($r^2 = 0.34$).

A comparison to two more recent studies confirms relationships that are similar to those observed with GL AoA. The Bristol Norm study (Stadthagen-Gonzalez and Davis, 2006) was performed in an identical way to the GL study and comparing the lists of acquisition ages results in a mean absolute error of approximately 0.5 which is much lower than the .9 to 1.5 relative to GL. The recent AoA list of Cortese and Khanna (2008) showed an increase in correlation relative to the GL study ($r^2 = 0.43$) potentially reflecting change in the acquisition process due to temporal effects.

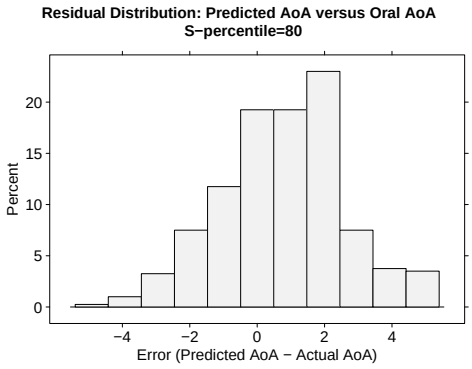

Figure 4: The difference distribution between the GL and the inferred AoA from Web 1-12 is skewed to the right as would be expected since written AoA is higher than oral AoA. Relaxing the definition of readability by decreasing s results in higher inferred acquisition ages. Values of s in $[0.5, 0.9]$ produced reasonable results, with $s = 0.65$ achieving smallest mean absolute error.

Those words that have the same written and verbal acquisition age are partially attributable to those words learned prior to first grade. Many words are learned between the ages of 2 and 5, while reading materials are typically not assigned a grade level of less than 1 or age 6. Approximately 40% of the words assigned the same grade level by both Gilhooly and our prediction had an AoA of 1st grade.

In some cases, the ages of acquisition obtained from readability data is actually lower than the ages reported in the older oral studies. This phenomenon is likely caused by a combination of a shift in educational standards, a change in social standards, or estimation errors due to sample size and modeling assumptions. Approximately

30 years have passed since Gilhooly and Logie's study was conducted. Specifically, society has made efforts to enhance the safety and health of children and to increase the attention to science education in very early grades. For example, the word `drug` appeared in writing 0.94 grades earlier than the age in which it was acquired orally according to the GL study. The newer Bristol Norm study confirms this observation as it predicts a decrease in grade level for `drug` of 0.88 over GL as well. A similar decrease in acquisition age relative to the GL norms was noted for many other words such as `hypothesis`, `conclusion`, `engineer`, `diet`, `exercise`, and `vitamin`.

3.3 Global Readability Prediction

Once acquisition age distributions are available, whether estimated statistically from data or obtained from a survey, they may be used to predict the grade level of novel documents. Specifically, the model predicts readability level t^* for a novel document d if it is the minimal grade for which readability is established:

$$t^* = \min\{t : P(d \text{ is readable at age } t) \geq \beta(t)\} \tag{5}$$

where $\beta(t)$ is a parameter describing the strictness of the readability requirement. Note that we allow $\beta(t)$ to vary as a function of time (grade level). We discuss the justification for this below.

A critical issue for reading difficulty prediction is how to handle words that appear in a new document that have never been seen in the training/development texts. In a statistical approach, the solution to this smoothing problem has two steps. First, we must decide how much total probability mass to allocate to all unknown words. Second, we must decide how to subdivide this total mass for individual words or classes of words using word-specific priors.

Our experience suggests that the first step of estimating total probability mass is particularly important: the likelihood of seeing an unknown word increases as a function of total vocabulary size, which is continuously growing with time. We model this by defining the following dynamic threshold

$$\beta(t) = \frac{\exp(at - 0.5)}{1 + \exp(at - 0.5)}. \tag{6}$$

We learn the growth rate parameter a in (6) from the data at the same time as we learn the readability model's quantile parameters $s = 1 - \epsilon_1$, $r = 1 - \epsilon_2$. The range of the resulting $\beta(t)$ is typically 0.5 in lower grades, increasing to 0.9 in higher grades. We discuss fitting these parameters and their optimal values further in Sec. 3.3.1. We found that using any fixed β value for all grades was generally much less effective than a dynamic $\beta(t)$ threshold, and so we focus on the latter in our evaluation.

For the second (word-specific) smoothing step, we simply assign uniform probability across grades, once the total unseen mass is determined. More sophisticated word-specific priors incorporating word length, morphological features, semantic clusters and so on are certainly possible and an interesting direction for future work.

In the following section we conduct three experiments involving readability prediction. First, we confirm the effectiveness of the AoA-based model compared to other predictive models. Second, we examine how prediction effectiveness is affected when our learned (written) acquisition ages are replaced with existing oral AoA norms. Third, we examine the ability of our model to generalize to new content by training and testing on different (non-overlapping) corpora.

3.3.1 Effectiveness of Readability Prediction

In order to assess the effectiveness of our model in predicting the readability grade levels of novel documents we apply the model to two corpora. First, we use the Web 1-12 corpus to learn optimal parameter values for a , r, and s and then assess prediction error using a test-training paradigm for the proposed model, Naive Bayes, and support vector regression. Second, the trained model is applied with to the Reader A-Z corpus and the results are compared with alternative semantic variables. Because corpora can vary significantly in text homogeneity, amount of noise, document size, and other factors, training and testing across different corpora – rather than relying on cross-validation with a single pooled dataset – gives valuable information about how a prediction method might be expected to perform on data with widely different characteristics. This particular choice of Web 1-12 for training and ReadingA-Z for testing was arbitrary.

To evaluate the best values for the a parameter in (6) and s, r parameters in Definition 1 we gen-

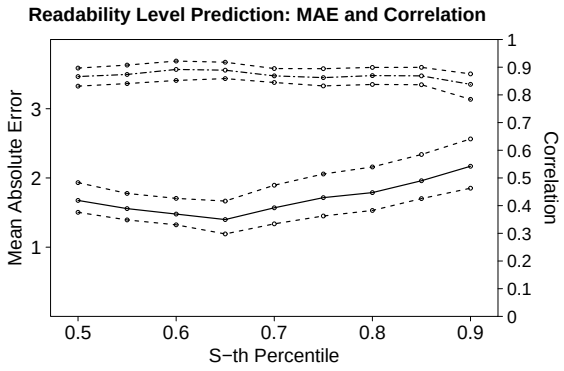

Figure 5: Mean absolute error (MAE) and correlation coefficient as functions of the quantile parameter s at optimal levels of a and r, averaged over 100 training/test samples. The MAE is displayed as the solid line and is aligned with the left axis while the correlation is displayed as a dashed line and is aligned with the right axis. 90% bootstrap confidence intervals are displayed.

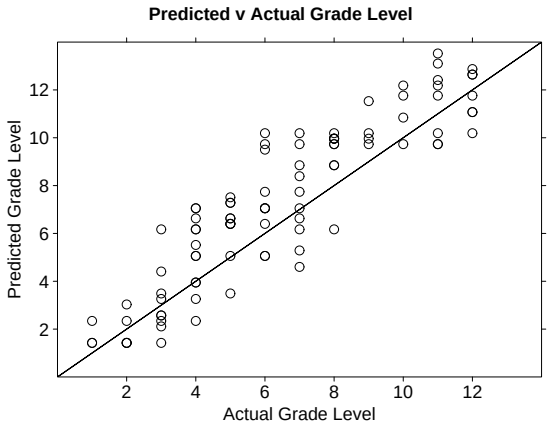

Figure 6: The scatter plot demonstrates the strong relationship between predicted and actual global readability levels.

Prediction Rule	MAE	LB	UB
Age of Acquisition	1.40	1.19	1.67
Naive Bayes	1.98	1.71	2.26
SVR (word frequency)	1.86	1.69	2.06
SVR (AoA percentiles)	1.36	1.22	1.58
Grade 6	2.92	-	-

Figure 7: 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.

erated 100 independent test and training samples and computed the mean absolute prediction error (MAE) and the correlation coefficient between the predicted and actual levels. Figure 5 (left) shows these two quantities: in each group of three lines, the top and bottom lines delineate the upper and lower 90% confidence bounds for the middle line. Each middle line gives mean error or correlation as a function of the quantile parameter s at optimal levels of r and a, averaged over the 100 training/test samples. The optimal value of s for both quantities is around 0.6 (0.65 for the MAE). The optimal value for parameter a was approximately 1.55. The best MAE is 1.4 which compares favorably to the 2.92 MAE obtained by always predicting Grade 6 which is the optimal "dumb" classifier in the sense that of all constant predictors it provides the smallest expected MSE over a uniform grade distribution as is the case with the Web1-12 corpus. Figure 6 is a scatter plot comparing predicted grades vs. actual grades, with a strong correlation of 0.89.

We compared the predictions of model (3) to two standard classifiers: naive Bayes and support vector regression (SVR). SVR was applied twice using different sets of features - once with the document word frequencies and once with the estimated AoA percentiles for the document words. The document word frequency vector is comparable to the semantic component of the machine learning approach used by (Heilman et al., 2008). The 75-25 training-test model building paradigm was used over documents from grades 1 to 12 to obtain predicted values. The MAE for these predictors and their 90% confidence intervals are shown in Figure 7. Predicting readability using word frequencies had inferior performance, with the naive Bayes model performing poorly and the SVR and Rasch model obtaining MAE around 1.4.

In the second experiment, we compared our model to published correlation results (Collins-Thompson and Callan, 2005) for multiple alternative semantic variables using the same Reading A-Z corpus, with the results shown in Fig. 8. Details on these semantic variables, which have been used in previous statistical learning approaches, are available in the same study. Interestingly, the correlation of the model was comparable to ex-

	Correlation		Correlation
GL (Web)	.65	UNK	.78
GL (WR)	.40	Type	.86
Bristol (Web)	.76	MLF	.49
Bristol (WR)	.57	FK	.30
Inferred (Web)	.59	Unigram	.63

Figure 8: Comparison of the correlation of AoA and other semantic variables with grade level for the Reading A-Z corpus, showing the AoA model with the dynamic threshold compares well to existing methods. The competitor methods used are from (Collins-Thompson and Callan, 2005) and comprise the Smoothed Unigram, UNK (relative to revised Dale-Chall), TYPE (number of unique words), MLF (mean log frequency), and FK (Flesch-Kincaid readability).

isting variables, but did vary depending upon the source of AoA. Note that because the Reading A-Z texts were assigned grades by their creators using some of the same semantic variables (e.g. Type), it is not surprising that those variables perform especially well on this dataset.

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 Rasch 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 Rasch 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 Figure 7 show that SVR and the dynamic threshold prediction rule perform similarly well, suggesting that Definition 1 and the Rasch model are suitable models for readability prediction.

3.3.2 Prediction with Existing Acquisition Age Norms

We now examine how predicting readability of novel documents using acquisition ages obtained in surveys perform in comparison to the ages obtained from the maximum likelihood estimation.

We use the GL and Bristol age of acquisition norms. The intersection of AoA norm data and the Web Corpus are 1217 and 1012 words respectively for the GL and Bristol measure; additionally, the highest grade level associated with these word sets

Prediction Rule	S-th Percentile	Dynamic Threshold
Age of Acquisition	1.69	1.40
GL Norms	1.73	1.42
Bristol Norms	1.97	1.79

Figure 9: The Gilhooly and Logie AoA norms and the Bristol norms are independent sources for ages of acquisition. A comparison of the prediction quality using these norms shows two things: 1) the definition provides comparable prediction quality using expert norms, and 2) the dynamic threshold $\beta(t)$ improves prediction over the static threshold (optimal s-th percentile) for the norms.

AoA Source	Web 1-12	Weekly Reader
Inferred (Weekly Reader)	-	.91
Inferred (Web 1-12)	1.89	-
GL	2.05	1.14
Bristol	1.57	1.34

Figure 10: The readability of WR documents was predicted using 4 sources of AoA data. The parameters of the prediction model were fit using only the Web data, or the WR data, or both sources in the case of the GL and Bristol norms AoA data.

are eight and seven respectively. 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. Interestingly, the optimal ranges of s-percentile, from 92 to 100, were the same for both the GL and Bristol norms. Table 9 shows that the prediction accuracy obtained using the GL Norms was almost identical to that obtained with the inferred AoA, while the Bristol Norms performed as well as some of the competitor procedures.

3.3.3 Prediction Effectiveness across Different Corpora

To provide additional evidence for our model's ability to generalize to new corpora, we examine how the learned r and s values vary when the model is learned on one corpus and evaluated on another, and how this affects the accuracy of the readability prediction.

Figure 10 demonstrates the corpus used for tuning the readability prediction has a large impact on the quality of the prediction. Comparing the MAE of the readability predictions on WR data

when the age of acquisition is inferred from Web data to the MAE when the AoA is inferred from WR data shows the error rate more than doubles from 0.90 to 1.89. The increase in error rate also appears when the age of acquisition for WR data is predicted using the AoA norm data. In this case the prediction was performed using the parameters identified when the model was trained on Web data and when the model was trained on WR data. In each case a tendency to overfit appears as the MAE increases from 1.14 to 2.05 for the GL norms and 1.34 to 1.57 for the Bristol norms. Interestingly, the Bristol norms perform better on WR data when fit using the Web data, while the GL norms perform better when fit using the WR data.

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; 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 related study is that by Crossley et al. (2007) who used AoA to help discriminate between authentic and simplified texts for second-language readers.

In the past decade, 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 et al., 2008), k-NN classifiers and combining semantic and grammatical features (Heilman et al., 2007).

The growing number of features investigated by these machine learning approaches reflect the fact that reading difficulty is a complex phenomenon involving many factors, from semantic difficulty (vocabulary) to syntax and discourse complexity, reader background, and others. While a full-featured comparison between previous approaches that includes AoA features would be very interesting, our goal in this study was to provide a clear analysis of the most fundamental factor of readability, semantic difficulty, which accounts for 80-90% of the variance in readability prediction scores (Chall and Dale, 1995). Because AoA is a semantic, vocabulary-based representation, we compare its effectiveness with the corresponding *semantic components* from previous machine-learning approaches in Sec. 3.3.1.

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 e.g., (Chall and Dale, 1995) and states that document readability level is determined by most people understanding most words.

The connection between word acquisition and readability is both intuitive and useful. It allows two degrees of freedom $s = 1 - \epsilon_1$ and $r = 1 - \epsilon_2$ 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. Experimental results show that the proposed model is also effective in terms of predicting readability level of documents on multiple datasets. It compares favorably to naive Bayes and support vector regression, the latter being one of the strongest regression baselines.

Acknowledgments

The authors thank Joshua Dillon for downloading the weekly reader data and pre-processing it. The work described in this paper was funded in part by NSF grant DMS-0604486.

References

J. S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Brookline, MA.

K. Collins-Thompson and J. Callan. 2005. Predicting reading difficulty with statistical language models. *J. of the American Soc. for Info. Science and Tech.*, 56(13):598–605.

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

M. Cortese and M. Khanna. 2008. Age acquisition ratings for 3000 monosyllabic words. *Behavior Research Methods*, 40:791–794.

S. A. Crossley, P. M. McCarthy, and D. S. McNamara. 2007. Discriminating between second language learning text-types. In *Proc. of the Twentieth International Florida Artificial Intelligence Research Society Conference*.

H. A. David and H. N. Nagaraja. 2003. *Order Statistics*. Wiley, Marblehead, MA.

E. Fry. 1990. A readability formula for short passages. *Journal of Reading*.

K. J. Gilhooly and R. H. Logie. 1980. Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation*, 12:395–427.

M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proc. of the Human Language Technology Conference*.

M. Heilman, K. Collins-Thompson, and M. Eskenazi. 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*.

S. E. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proc. of the Association of Computational Linguistics*.

H. Stadthagen-Gonzalez and C. J. Davis. 2006. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38:598–605.

A. J. Stenner. 1996. *Measuring reading comprehension with the Lexile Framework*. Metametrics, Inc., Durham, NC.

J. D. Zevin and M. S. Seidenberg. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, 47(1):1–29.