This is the accepted version of the manuscript

Kapelner, A., Soterwood, J., NessAiver, S., & Adlof, S.M. (2018). Predicting contextual informativeness for vocabulary learning. *IEEE Transactions on Learning Technologies, 11*, 13-26. doi:0.1109/TLT.2018.2789900.

Published Jan-March, 2018

Predicting Contextual Informativeness for Vocabulary Learning

Adam Kapelner, Jeanine Soterwood, Shalev Nessaiver, Suzanne Adlof

Abstract—Vocabulary knowledge is essential to educational progress. High quality vocabulary instruction requires supportive contextual examples to teach word meaning and proper usage. Identifying such contexts by hand for a large number of words can be difficult. In this work, we take a statistical learning approach to engineer a system that predicts informativeness of a context for target words that span the range of difficulty from middle school to college level. Our database (released open source) includes 1,000 hand-selected words associated with approximately 70,000 contextual examples gathered from the Internet. Our training data included each context rated by 10 individuals on a four-point informativeness scale. We decompose the text of each context into a novel collection of approximately 600 numerical features that captures diverse linguistic information. We then fit a nonparametric regression model using Random Forests and compute out-of-sample prediction performance using cross-validation. Our system performs well enough that it can replace a human judge: for a target word not found in our dataset, we can provide curated contexts to a student learner such that most of the contexts (54%) feature rich contextual clues and confusing contexts are rare (<1%). The quality of our curated contexts was validated by an independent panel of high school language arts teachers.

Index Terms—Adaptive and intelligent educational systems, statistical software, machine learning, text mining, language summarization

1 Introduction

Knowledge of words, and their meanings, is essential for spoken and written communication. Measures of vocabulary knowledge predict literacy, academic, and cognitive outcomes throughout one's lifespan [1], [2], [3], [4] and mastery of core vocabulary is necessary for developing adult reading comprehension skills [5]. Improving vocabulary knowledge is an important goal for many adolescent students, including those who struggle academically [6] as well as those preparing for college entrance exams.

How does one go about increasing vocabulary knowledge? For a first language, it is well established that most words are learned via context during normal reading [7], [8], [9], [10], [11], [12], [13], [14] (for a summary of the causal mechanisms and assumptions, see the introduction section of [15]). Hence the obvious approach to increasing vocabulary is to provide the reader with many contextual experiences. This is best done via wide reading as printed text provides exposure to a wider range of words than conversation [13], [16], and there is an unequivocally strong relationship between reading frequency and vocabulary size in both children and adults [17], [18], [19]. However, some have argued that wide reading is not only inefficient but also disadvantages individuals with reading difficulties [20]. Thus, direct instruction of carefully selected, high utility words is also a recommended strategy. Systematic reviews of vocabulary intervention studies (most of which targeted younger readers) recommend teaching word meanings through a combination of dictionary definitions with examples of word usage in context [21], [22]. Cumulative experiences with words in varying contexts enable learners to abstract from specific episodes [23], [24] and thereby acquire rich representations of word meanings.

The problem of efficient vocabulary teaching is ripe for an instructional technology solution. Toward this end, we are developing and evaluating "DictionarySquared", an Internet-based vocabulary software intervention that provides targeted instruction of high utility words using authentic contextual examples. The DictionarySquared system also allows for randomized controlled experiments that investigate aspects of vocabulary learning. Both uses are described in detail in Adlof et al. [25].

The purpose of this paper is to describe a system that automatically identifies informative contextual examples for first language vocabulary instruction for high school students. (Note that the effectiveness of second language vocabulary acquisition via context during wide reading is not well-established [15]). Although this constitutes a core piece of instructional technology within DictionarySquared, our technology presented herein can be of general interest to anyone who wishes to curate informative contexts for word learning, including second language practitioners and researchers.

1.1 The Importance of Informative Contexts

As Beck et al. [26] explain, "although it may be true that the learning of new words is facilitated by some contexts, it

- *Adam Kapelner is an Assistant Professor at Queens College, CUNY, Department of Mathematics, Kiely Hall 604, 65-30 Kissena Blvd., Queens, NY 11367*
  E-mail: kapelner@qc.cuny.edu
- *Jeanine Soterwood is principal at Littleforest Consulting, LLC, 680 Bucher Ave, Santa Clara, CA 95051*
- *Suzanne Adlof is an Assistant Professor at the University of South Carolina, Department of Communication Sciences and Disorders, Keenan Building, Room 330, Columbia, SC 29201*
  E-mail: sadlof@mailbox.sc.edu

Manuscript xx xx, xxxx; revised xx xx, xxxx. All authors contributed to the conceptualization of this work; Adlof supervised collection of informativeness ratings; Kapeler, NessAiver and Soterwood generated text features; Kapelner built machine learning models; Kapelner and Adlof drafted, edited and revised the manuscript.

is not true that every context is an appropriate or effective instructional means for vocabulary development." Beck et al. [26] describe four categories of contexts: misdirective, non-directive, general and directive. Misdirective contexts "direct the student to an incorrect meaning" which is harmful when learning a word initially. Nondirective contexts lack contextual clues and thereby provide no assistance. General contexts give enough clues for the learner to frame the word into a general category. Directive contexts are full of rich cues and are thereby the most pedagogically effective.

Contextual word learning experiments have demonstrated that the informativeness of instructional contexts impacts learning. Frishkoff et al. [11] presented adults with target words in six sentence contexts. The informativeness of the contexts was systematically manipulated, such that some words were presented in six directive contexts, some words were presented in five directive and one misdirective contexts, and some words were presented in three directive and three misdirective contexts. Directive contexts were constraining for the target word, whereas misdirective contexts were constraining for a distractor word that had a different meaning from the target word but overlapped in phonology and orthography. Participants were asked to generate a succinct definition for the target word after each contextual experience. Analyses examining the accuracy of definitions generated across trials revealed a significant interaction between trial and context quality. Follow up analyses of the generated definitions indicated no differences in performance for words taught with five vs. six directive contexts; however, performance for words taught with three directive and three misdirective contexts was significantly worse than in the other two conditions. Analyses of synonym judgment accuracy at pre- and post-test also revealed a significant interaction between test time and context informativeness, such that gains in accuracy were significantly greater for words taught with more directive contexts.

Another study presented adults with six contexts for each target word, with context varying in their informativeness [27]. This time, no contexts were misdirective; instead, some words were taught in six directive contexts, some words were taught in six nondirective contexts, and some words were taught with half directive and half nondirective contexts. Results were similar to Frishkoff et al. [11]. Analyses of generated definitions revealed a significant interaction between trial number and informativeness, such that the accuracy of definitions generated for words presented in all nondirective contexts did not improve over trials, but the accuracy of definitions for words presented in half or all directive contexts significantly improved. Also, analysis of a synonym judgment task administered at pre- and post-test revealed significant interactions between test time and context informativeness, such that words taught in half- or all-directive contexts were learned better than words taught in all nondirective contexts.

Adlof et al. [28] also examined contextual word learning by children (aged 9-12 years) and adults. In their studies, some words were taught in two or three directive contexts, following zero, one, or four familiarization exposures in nondirective contexts. A matched set of "control" words were included on pre- and post-tests, but never appeared in context. Analyses of synonym generation and synonym matching tasks administered at pre- and post-test revealed significant differences between taught words and control words, but no significant differences between the words that were pre-familiarized in nondirective vs. not familiarized.

Taken together, these results suggest that in designing an optimal vocabulary tutor, the majority of contexts should be directive (or at least generally directive). Misdirective contexts need to be eliminated as best as possible as they detract from learning [11]. As for the ambiguous, nondirective contexts, it appears they don't facilitate learning by themselves, and they neither hurt nor help vocabulary learning, once supportive contexts have been provided [27]. Thus, the purpose of this paper is to describe a system that automatically identifies contextual examples with roughly the informativeness distribution outlined above. This system will help to optimize the selection of contexts for our DictionarySquared program, but can also be used by teachers or others who wish to efficiently identify example contexts for vocabulary instruction.

1.2 The Specifics of our Problem

We wish to engineer a system that takes as an input a target word (a single word that we wish to teach a student) and a context (a block of text containing that word) and output a binary decision: "use" or "not use" in the DictionarySquared vocabulary-teaching system.

We limit the scope of acceptable words to be from approximately late middle school level (e.g. "relevant", "ethnic", "promote"), late high school level (e.g. "surly", "vestige", "primordial") and up to college level and beyond (e.g. "meretricious", "vitiate", "bucolic"). The system should be able to take, as an input, any word intended for vocabulary study in this range with an associated written context containing the word.

DictionarySquared teaches by showing contexts whose format is mostly textual. Its strategy is focused on students reading many short contexts instead of a few long contexts (see [25] for specifics). Since our system presented here will be a core technology within Dictionarysquared, we limit what we mean by a "context" to blocks of text between 42–65 words (on average, 54.2 ± 3.4) where each features the target word (in the target inflection) at least once, where virtually all of the contexts (a) begin at the beginning of a sentence, (b) end at the end of a sentence and (c) do not span between paragraphs of the source text it was excerpted from.

As an example, consider the target word "proclivity" (in bold). We would like to select informative contexts such as:

> Some people have a genuine **proclivity** for motion sickness and will undoubtedly suffer more during rough seas. According to medical professionals, seasickness is more prevalent in children and women. On the other hand, children under 2 seem to be immune from the ailment. Of equally interesting note, elderly people are less susceptible ... [29, third paragraph]

that is likely between general and directive, and discard uninformative contexts such as:

> Yet more additions to the links bar, and kind of by way of addressing the massive **proclivity** of "Da Phenomenon," I'd like to shower a little love on some of my favourite bloggers, be a little selective for a change. Somewhat ruefully I was reflecting that doing one of my pictorial bloggers break-outs again is now becoming increasingly unlikely. [30]

that is likely between misdirective and non-directive.

Within the scope of our problem, we identify two slightly different challenges. The first is to classify the informativeness of a new, never previously seen target word embedded in a new, never previously seen context. The second is to perform the same classification on a target word seen before but a context never previously seen. We will refer to these two similar challenges for the duration of this paper as [word unseen] and [word seen] respectively.

1.3 Previous Research

Finding examples of words to promote learning has been considered before. Brown and Eskenazi [31] developed "REAP", a search engine for contextual examples optimal for personalized vocabulary learning by considering students' grade level defined by word histograms [32] while keeping track of words previously seen. We share the final end-goal to foster personalized vocabulary growth, but we focus on providing rich contexts for individual words only. We also make implicit use of their student leveling by including features based on word frequencies and n-grams in our prediction models. Mostow et al. [33] developed a method to create short (about 6-words) informative contexts containing a single sense of a polysemous word using n-grams. Here we focus on informativeness without regards to the word sense. We also make implicit use of the n-grams features developed therein. Hassan and Mihalcea [34] automatically classified entire documents as "learning objects" [35] or "learning assets" for concepts and showcased their system on fourteen computer science concepts. Although their goal was quite different than teaching vocabulary, we make use of their method of hand-engineering features and employing supervised machine learning to evaluate educational value.

Mostow et al. [36] studied the prediction of informativeness for individual contexts using a subset of 13,000 contexts for easy and medium difficulty words in our database. Their preliminary analysis indicated that a linear model fit with measures of context length, context word frequency, context readability, local predictability (derived using Google n-grams), co-occurrence and distributional similarity as covariates predicted informativeness better than chance. However, the binary classification of good vs. bad contexts was insufficient to generate a set that included mostly good contexts. Even using the most stringent predicted probability rating (which discarded over 95% of the contexts) resulted in the acceptance of more bad contexts than good contexts. In this study, we make use of the full set of available data, expand the features to include more sophisticated n-grams and semantic similarity features as well as a large number of natural language processing indices, and we use advanced machine learning methods that lead to better performance. Additionally, our system's framework, model assumptions and presentation differ considerably.

2 Methods

To make a prediction of whether or not a context is to be used or not used in a vocabulary learning system, we employ supervised statistical learning [37, Chapter 2]. To do so, we require two things: (1) training data, which includes both (a) informativeness ratings for each contextual example and (b) numerical "features" of the contextual examples that correlate with the informativeness ratings, and (2) a statistical model and a means to fit this model using the training data. We explain these two requirements below and discuss how the statistical model's performance is assessed.

2.1 Training Data

2.1.1 Target Words and Associated Contexts

The current DictionarySquared system teaches 1,000 words split into 10 "bands" of 100 words each. These words span a range of difficulty to meet the needs of low, average, and high skilled high school students. To curate our list, we began with over 2,500 words compiled from Coxhead's Academic Word List [38] as well as lists of suggested words to study in preparation for standardized tests such as the ACT, SAT, and GRE. We derived estimates of word difficulty based on word frequency and dispersion norms [39] and age-of-acquisition estimates [40]. These two indices were highly, but not perfectly correlated within the list ($r = +0.77$). We then divided the list into 10 difficulty bands and hand-selected 100 words in each band that would presumably be considered "useful" for instruction according to criteria described by vocabulary experts. These included words that: characterize written text and are general enough to be found across academic content domains [41], [42], might be difficult to learn from everyday conversation, but occur frequently enough in academic texts to be of assistance to the comprehension process [41] and further, those that are generally explainable using familiar concepts [43].

For each word, we query the DictionarySquared database for contexts that contain the word in the same inflection. This DictionarySquared corpus was populated between 2008-2010 using the Google Web API (since defunct). Contexts came from text separated by spaces within one html tag (to enforce contiguous text) and result order was randomized. Care was taken to drop contexts with duplicate sentences (as defined by sharing one complete sentence) to ensure uniqueness of each context. Note that unlike a Google News search, search results from the web API were not clustered into similar items. The contexts are devoid of illegal characters or inappropriate words (according to a handmade list of ≈900 words) and do not have too many non-letter characters.

We then pruned the original 1,000 word list ensuring each word has more than 20 associated contexts (on average,

the words have 72.7 $\pm$ 20.7 contexts) for a total of 933 words with 67,833 associated contexts. Thus, our training data frame has 67,833 individual rows which constitute training data examples for use in training and validating our statistical model.

2.1.2 Informativeness Ratings

After contexts were queried, each context was hand-rated using Amazon's Mechanical Turk (MTurk), a world-wide market for one-off concise tasks that has been validated to be accurate for natural language tasks [44]. Ten unique MTurk workers rated each context for a total of approximately 700,000 ratings. The contexts were rated on the same ordinal scale of "informativeness" pictured below:

- **Very Helpful**. After reading the context, a student will have a very good idea of what this word means.
- **Somewhat Helpful**.
- **Neutral**. The context neither helps nor hinders a student's understanding of the word's meaning.
- **Bad**. This context is misleading, too difficult, or otherwise inappropriate.

The above scale choices roughly correspond to the Beck et al. [26] scale without the use of the technical terms found therein. We numerically code the levels in our ordinal scale as +2, +1, 0, -1 respectively, i.e. the conventional default of even spacing. We feel it is appropriate to code neutral as 0, bad as negative and helpful as positive but the values -1, +1 and +2 are arbitrary. Future work can explore other encodings.

Several steps were taken in an effort to ensure quality of the collected MTurk ratings in accordance with methods and recommendations from past studies [45], [46], [47], [48]. First, raters were required to initially pass a qualification test which included contexts with known ratings. Second, individual rater agreement was monitored over time; raters whose response patterns appeared random or who otherwise showed substantial deviations from the crowd were disqualified from future rating assignments. This is similar in spirit to more advanced algorithms [49, for example]. Lastly, we included random "attention checks"; these consisted of eliciting a rating for clearly helpful and clearly unhelpful contexts from the more difficult bands (7–10) as determined by unanimous agreement from trained research assistants. Raters who frequently made errors on the attention checks were disqualified from future rating tasks, a strategy similar to the work of Oppenheimer et al. [50]. Even though we took principled steps, it was still difficult to root out all the low quality ratings. Cleaning out low quality ratings using principled methods [51, for example] is a worthy enterprise, but it is left to future work.

We then employ the sample average of the 10 ratings as the gold-standard label for each context. Previous research has found that averaging the ratings from 9 or more non-expert raters reached agreement with expert raters [44], [52]. We henceforth denote the average rating as y, indicating this is our dependent variable going forward. Note that the quality of the gold-standard y (along with our modeling) is ultimately validated by the independent teacher validation experiment that we describe in Section 4.

The density of our context ratings as well as a breakdown by band are pictured in Figure 1. Note the rating distribution is approximately Gaussian with an average of 0.59 ± 0.53.[1] These contexts make a good training data set for developing the system we wish to build, as the examples are drawn from a wide distribution of informativeness.

The fraction of misdirective contexts is at most 15% (those rated less than 0) and the fraction of directive texts is at most 19% (those rated greater than 1). The average context in our current corpus is therefore between nondirective and general. We can appreciate that the raw database has quite informative contexts as it was culled from reliable text sources. Thus, our job to curate a subset of *truly excellent* contexts is very challenging, a point we will return to in Section 4.

2.1.3 Feature Extraction

Applying learning algorithms to data requires a transformation from the raw data to a collection of features since the raw data representation (the text itself) is a suboptimal representation of the underlying information in the text [54]. Put another way, the raw characters of the text considered alone have negligible correlation with the informativeness rating.

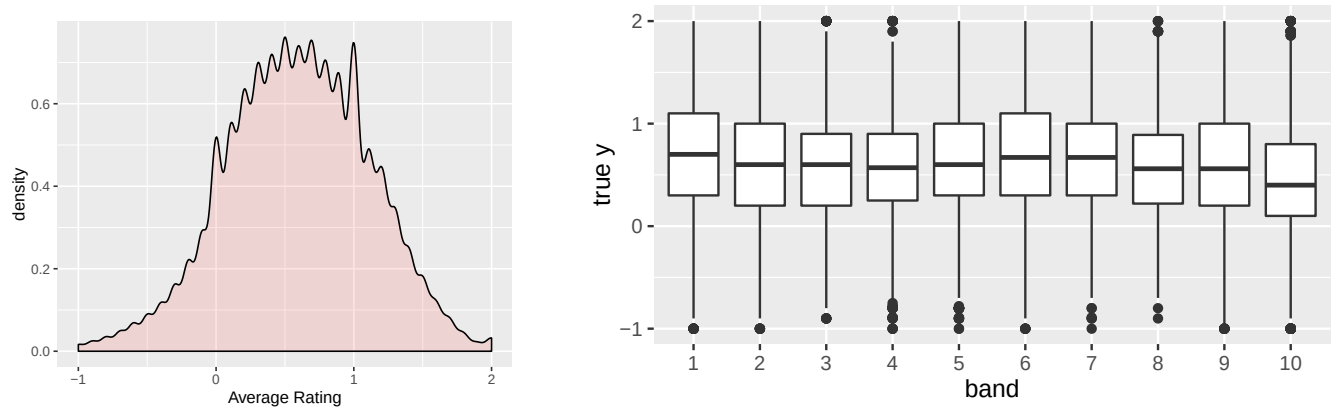
Thus, the next task consists of creating our own data representation — mapping the characters in the contextual examples to numeric features, i.e. functions g_j : text $\to \mathbb{R}$, which is a type of "text mining" [55], one of the goals of natural language processing (NLP). Then we will extract patterns from these numeric features, relating them to informativeness; this is a type of "machine learning".

The literature on data representation via textual feature extraction is vast and it is difficult to know which features should be extracted. We began with simple features such as number of sentences, words and punctuation counts, etc. We then read many hundreds of contexts ourselves and tried to conceptually isolate common attributes observed among highly directive contexts and other attributes observed in misdirective contexts. Through doing so, we tried to identify useful simplifying explanations or abstractions that helped to make sense of the rich data that is natural text; this is a form of disentangling features with correlation from ones without [56, Chapter 1].

We first observed that directive contexts have common expressions and phrases that use the word. This can be captured if we can find all such phrases for all words. Thus, we employ the Google Book Corpus including text from 1800-2000 [57] to calculate the number of times the target word is found in every configuration of n-grams up to 5-grams also including blanks (or stop characters). In addition to the probability estimate used by Mostow et al. [36], we include the raw count. For instance, for the word *defiant*, we have features for configurations found in Figure 2.

We also observed that directive contexts contain synonymous words (i.e. semantically related) as well as words that

1. Note that it is likely not valid to generalize this observation, as this dataset was collected from the Internet in quite an arbitrary way, and scored using a rating system that does not exactly reflect Beck et al.'s [26] categorization.



(a) Overall (b) By Band

Fig. 1: The distribution of y, the informativeness metric in our training set. All plots are generated with the `R` package `ggplot2` [53].

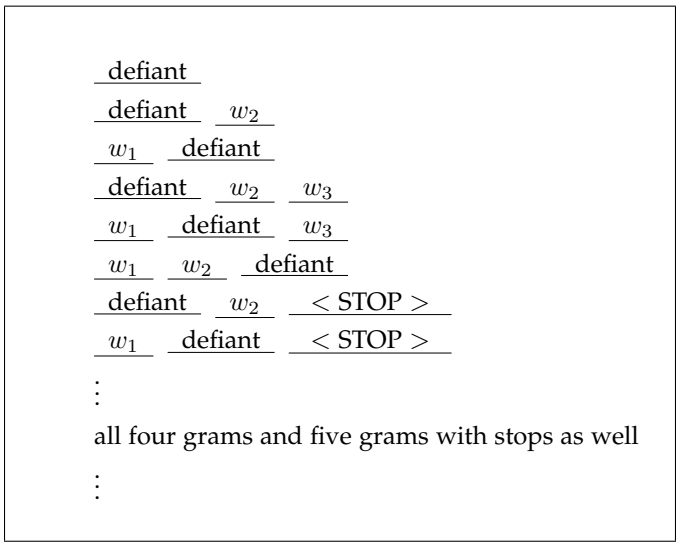


Fig. 2: An illustration of the n-gram features built for the word "defiant". w_i represents the surrounding words appeared in the context. If the target word appeared multiple times, we used the greater valued features.

frequently appear with the target words (i.e. collocated). We used the DISCO tool, which computes the top 200 most semantically related and collocated words for a target word [58]. Using collocations gathered from unannotated data as features is similar to the work by Yarowsky [59]. DISCO was run atop Davies' [60] Corpus of Contemporary American English corpus (COCA, 520 million words of text from 1990-2015 equally divided among spoken, fiction, popular magazines, newspapers, and academic texts) which confers significant advantage over DISCO's default Wikipedia corpus (as employed in [36]). We bin these 200 words into the top 10, the top 11–20, the top 21–50, 51–100 and 101–200. We then count common word stems that appear in the context within each bin and these counts constitute features. Note that the DISCO method is one of many methods to compute semantic similarity.

We then make use of recently developed NLP indices. First, we use features from the Tool for the Automatic Analysis of Lexical Sophistication [61], which calculates scores for 135 classic and newly developed lexical indices related to word frequency, range, bigram and trigram frequency, academic language, concreteness-abstractness (a metric known to be related to vocabulary learning, see [62]) and psycholinguistic word information. Second, we use features from the Tool for the Automatic Analysis of Cohesion [63] which calculates scores for 150 classic and recently developed indices related to text cohesion, work closely related to the development of "Coh-Metrix" [64]. Third, we use a subset of the Sentiment Analysis and Cognition Engine [65] which reports on over 3,000 classic and of lexical categories and 20 component scores related to sentiment, social cognition, and social order.

In total, we computed 615 features that attempt to span the high-dimensional space of linguistic information and thus may be predictive of informativeness. For a complete list of the features, see Appendix C. Note that the target word *itself*, a categorical feature, is not one of the 615 as this would not be generalizable nor helpful for building the [word unseen] system. Also note that our feature extraction strategy leverages data outside of our collected training data set via the use of these corpuses and NLP indices, a strategy that has been coined "cotraining" [66].

With 615 features, it is possible that many are collinear as they express the same information about the text. To investigate this possibility of our feature set being "overspecified", we ran a principal components analysis (centered and then scaled) and plotted the cumulative sorted normed eigenvalues in Figure 3. Each of normed eigenvalues captures the percentage of the variance explained in each principal component; the cumulative sum of these normed eigenvalues captures the percentage of the variance explained cumulatively among the principal components.

We see from this plot that we can cut the dimension of the feature space by at most a factor of two and still retain nearly all of the information (95% of the variance is explained by 306 features). Using 300 features instead of the full 615 is not a significant compression of the data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2789900, IEEE Transactions on Learning Technologies

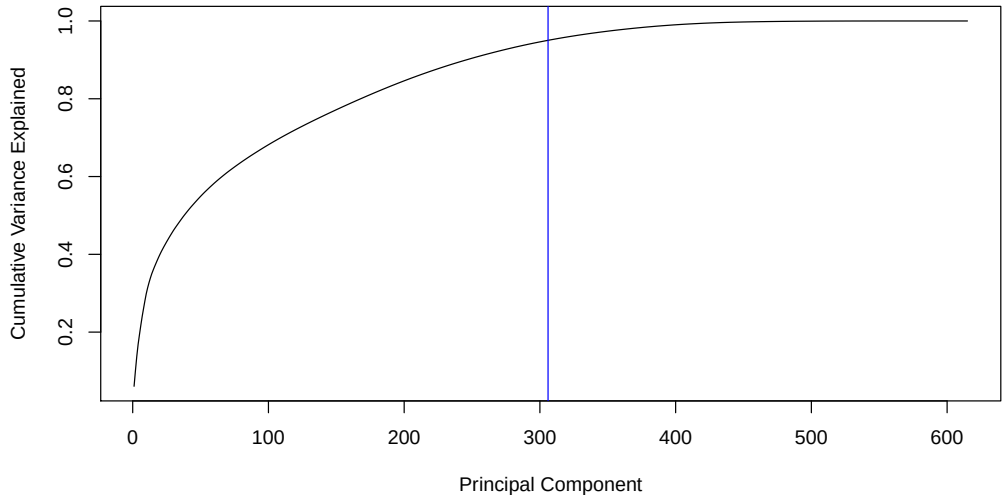


Fig. 3: The cumulative sum of sorted normed eigenvalues by principal component ordinal number. The vertical blue line illustrates the cumulative principal component number that explains 95% of the variance in our set of features.

during modeling; thus going forward we use the full feature set. This confers the additional advantage of allowing us to investigate individual variable importance directly in Section 3.4.

The full dataset, 70,000 $\times$ 615, the individual Turker ratings and the original text of the contexts is available and open source (see Section on "Replication").

2.2 The Statistical Model and its Use in Prediction

We begin with the continuous response variable of the average informativeness rating (Section 2.1.2) but then must make a binary decision on new contexts to administer a context to the student (one) or not to administer a context (zero). This binary decision is based on costs which vary across the spectrum of the informativeness rating $[-1, +2]$. These costs are quite asymmetric as it costs more for a student to be mistakenly presented with a poor context with informativeness near -1 than it does for us to mistakenly reject a good context with informativeness near +2 (see Appendix A for a mathematical description). In short, we have a regression estimation problem where decisions are made with a threshold based upon entire distributions. Below, we will develop a heuristic validation scheme appropriate to this nontraditional problem.

We employ random forests (RF) [67] to model the average informativeness rating y as a function of the 615 features in the training data. RF is widely used for non-parametric estimation of continuous functions (regression) by flexibly fitting complex non-linearities and high-order interactions using its tree structure without overfitting [37, Chapter 15]. Predictions from this RF model will be continuous. In order to make the binary decision, we use a threshold which we denote $\hat{y}_0$ (see page 12 of [37]) i.e. if $y \geq \hat{y}_0$, we administer the context (1, use) and if $y < \hat{y}_0$ we do not administer the context (0, throwout). Thus, our choice of $\hat{y}_0$ is chosen to minimize the overall asymmetric costs (explained in the next section).

2.3 Performance Validation

We now describe performance assessment for both the [word unseen] and the [word seen] task. We first split the full training data into a model-building set and a holdout or "test" set. The model building set is used to fit an RF model. The test set is then predicted via the fitted RF model and the results are compared to its true y values to determine performance; this is the "out-of-sample" (oos) estimate or generalization estimate. Validations performed oos guarantees that our performance estimate is not being inflated by potentially dishonest in-sample overfitting.

Generally speaking, the holdout consists of a random sample of the rows of the training data. Under this random sampling, the oos estimate would correspond to an honest estimate of the performance of [word seen] task only. Why? The target words in the holdout are the same target words (more or less) as those in the model-building set. This simulates new contexts for words already in the word bank.

To estimate the performance of the more realistic [word unseen] system, we randomize the target words themselves and holdout the rows corresponding with 10% of the words.[2] Thus, when the RF model makes predictions oos, it is predicting for contextual examples containing target words never heretofore seen.

However, any single 10%-90% test-training split can over or underestimate performance due to chance. To mitigate this possibility, we employ 10-fold cross-validation [37, Chapter 7.10].

The typical metric to consider in continuous prediction is oos R^2 or RMSE. Since our primary goal is to build a vocabulary learning system of which predicting the continuous informativeness metric is only a necessary intermediate step, R^2 or RMSE are not meaningful performance metrics. Here, our performance metric should reflect a total

2. We employ 10% as this is a common practice. There is little statistical theory at this time that recommends a hold-out size for estimating model performance. The largest used in practice is 20% and the smallest used is one observation (the so-called "leave one out" cross validation procedure).

cost function (Appendix A, Equation 2) which varies with the prespecified $\hat{y}_0$ threshold. Different threshold values result in differential distributions of oos informativeness that the system declares usable for future student consumption (we call this the "use distribution") and unusable for student consumption (the "throwout distribution"). To compare the use and throwout distributions, we compute empirical quantiles. We deem three quantile-based slices of the use distribution most important to monitor during the design of a student learning system: $Y < 0$, the poor contexts, $Y \in [0, 0.5)$, the non-informative educationally-neutral contexts and $Y > 1$, highly informative contexts. For the throwout distribution, we tabulate how often we throwout the very best contexts ($Y > 1$). A *holistic* view of these quantiles informs our heuristic total cost performance metric, not R^2.

To understand more clearly our starting point, Table 1 displays these quantile-slices for the original, uncurated data by band. We can see informativeness varies significantly by band. The highest bands, indicating words representing more nuanced concepts, have a larger the proportion of misdirective and non-informative contexts, as well as a lower proportion of richly informative contexts. Thus, our model's curation task is much more difficult at higher bands, a point we return to later (especially in Figure 5).

As $\hat{y}_0$ increases, the system will be more and more conservative as to which contexts it deems useful. Thus we can decrease the proportion of contexts with $y < 0$ and with $y \in (0, 0.5]$ and increase our proportion of $y > 1$ contexts by raising $\hat{y}_0$. But there is a tradeoff: the cost is greater false negatives; the system will throw out many contexts that are good for student learning. If our pool of potential contexts is large (such as the Internet), this is not a problem except on the very rare words (which are rare even across the entire Internet).

3 Results

Here, we focus on results for the [word unseen] system and we discuss results for the [word seen] system in Appendix B.

3.1 RF Results for the [word unseen] System

We vary the $\hat{y}_0$ threshold in order to investigate the possible binary decision systems that can be created from our continuous RF model.[3] For each threshold, we compute the empirical cumulative probabilities for the important metrics of Table 1 as well as the throwout percentage and display the results in Table 2. Each row of Table 2 estimates future performance: the cost of misdirective contexts, the cost of uninformative contexts, the reward of directive contexts and the cost in throwout percentage. Each row provides multivariate performance metrics.

For our purpose, we stress that misdirective contexts are costly since their appearance in a vocabulary training program for a given word has the potential to confuse the student [11], [27]. Thus, we believe our target proportion should be about 1–2% corresponding to a 7.5-15.1x fold reduction over the original contextual examples from the Internet (the training data). The system that performs at this level has $\hat{y}_0$ cutoffs 0.845–0.895 (highlighted in Table 2). In this range, we have approximately 10% of the contexts as uninformative. Most importantly, between 47–54% of the contexts will be general and directive featuring rich contextual clues that are supportive of rapid vocabulary learning. Put another way, the ratio of directive to misdirective contexts (after discarding) is 25:1 – 54:1. The price we pay is 91.6–96.6% throwout of our potential pool of directive contexts. We address the implications of this high cost in the discussion section. For the model with threshold $\hat{y}_0 = 0.895$, we illustrate the future use distribution compared to the training data in Figure 4.

3.2 Linear Model Results for the [word unseen] System

To demonstrate the predictive advantage of RF, we fit a linear model (via ordinary least squares regression). We varied $\hat{y}_0$ similarly and attempted to display the results that exhibit the 1–2% misdirective context proportion in Table 3 as a comparison. However, the linear model was not able to perform at that error rate (without ≈100% throwout). At higher rates, such as 3%, other cutoff performance was similar with the RF implementation but the throwout rate was higher.

3.3 Differential Performance by Band of the [word unseen] System

Note that the results in Table 2 and the use distribution illustration in Figure 4 represent an average across all words (and bands). It is likely that the system will have better performance on contextual examples with target words from lower bands. This association is expected as there are both more misdirective contexts and less directive contexts as band increases (review the differential distributions in Table 1).

We illustrate throwout rate by band in Figure 5. Note here that the throwout rate increases as the word band increases. The system does not work beyond band 6. Band 7 has 99.2% throwout and band 10 has 99.6% throwout.

Therefore, we recomputed the main results for just bands 1–6 and display them in Table 4. Our metrics of interest in the usage distribution largely remain the same but throwout has significantly improved. It seems the model's solution to uncertainty of usability among contexts in bands 7-10 is simply to omit them.

3.4 Feature Importance

Which of the 615 features contribute to predictive performance in the [word unseen] model? We queried our RF model[4] for its variable importance data and we plot the mean decrease in accuracy as measured by out-of-bag increase in mean squared error in Figure 6 (see [67] for more information). This metric is analogous to effect size in a multivariate regression. It accounts for collinearity

3. Note that binary classification decisions are popularly made with the help of a Receiver Operator Curve or Detection Error Tradeoff plot [68] (which plots throwout versus "errors"). Here we have two full distributions with differential costs; thus, analogous visual aids are not practical nor appropriate.

4. We used the entire dataset, default hyperparameter values, $n = 10,000$ samples per tree and 500 trees. Fitting more trees or raising the subsampling would be more computationally burdensome.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2789900, IEEE Transactions on Learning Technologies

Cutoff	Description of contexts	Proportion (Count) by Band 1	2	3	4	5	6	7	8	9	10
$y < 0$	potentially misleading	0.14 (671)	0.18 (815)	0.17 (762)	0.15 (637)	0.13 (590)	0.13 (624)	0.10 (419)	0.12 (524)	0.17 (763)	0.25 (1120)
$y \in (0, 0.5]$	non-informative	0.25 (1470)	0.27 (1576)	0.31 (1941)	0.33 (2068)	0.31 (1912)	0.28 (2165)	0.32 (2176)	0.37 (2799)	0.33 (2035)	0.34 (2154)
$y > 1$	rich in contextual clues	0.26 (1664)	0.21 (1238)	0.16 (1046)	0.14 (901)	0.19 (1298)	0.25 (1968)	0.22 (1639)	0.15 (1168)	0.19 (1205)	0.12 (764)

TABLE 1: Metrics computed about the oos use distribution. The cutoffs are for the real y values (the average rating). We then show the differential cutoffs by word band. Note: the columns do not sum to 1 as the range $y \in (0.5, 1]$ is not shown.

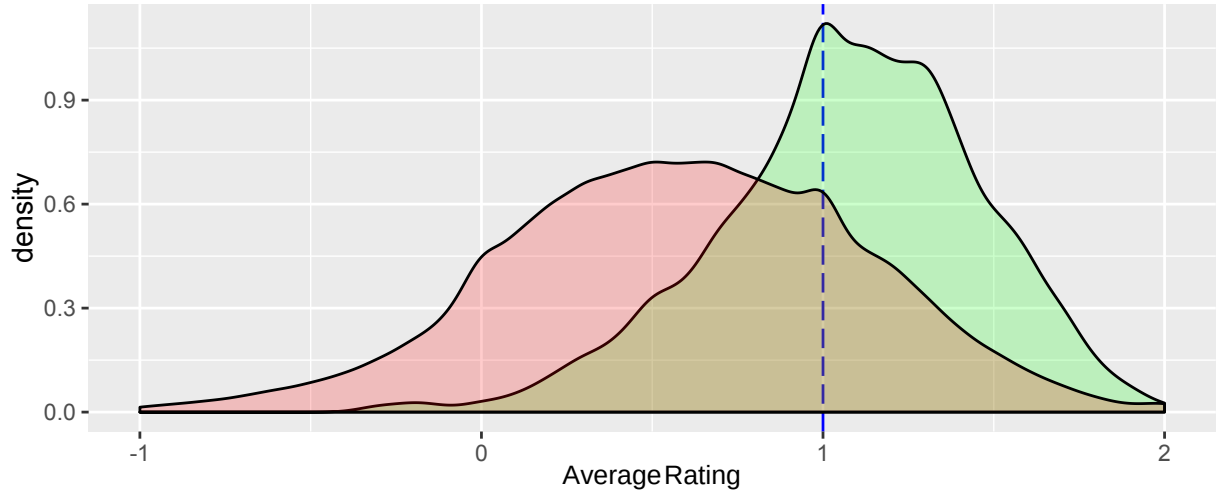


Fig. 4: The density of the informativeness use distribution (green) versus the training data distribution of informativeness (red) at $\hat{y}_0 = 0.895$. Contexts above the blue dotted line feature rich contextual clues that optimize student learning.

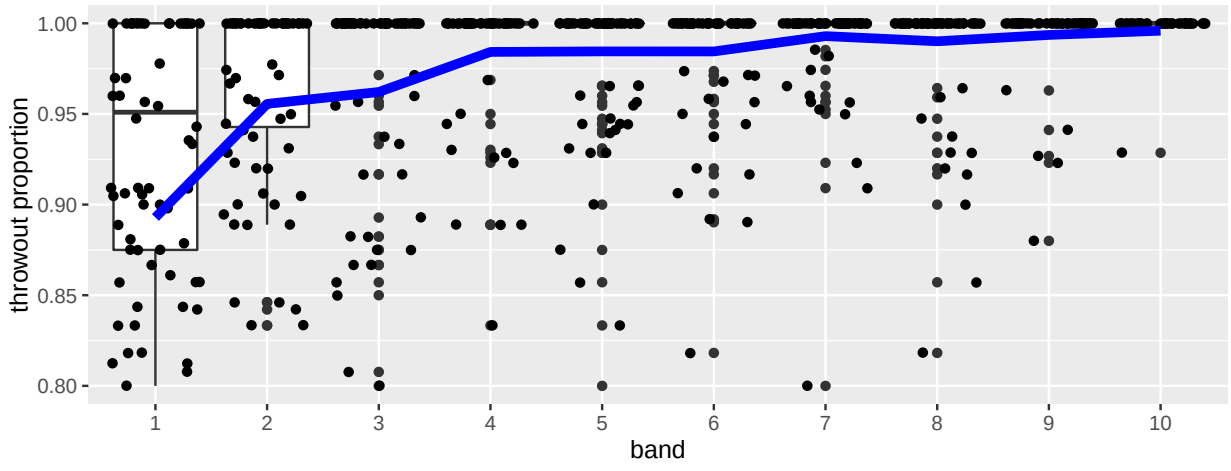


Fig. 5: Box-and-whisker plots of differential throwout by word organized by band for the RF model in the [word unseen] system at $\hat{y}_0 = 0.895$. The blue line plots the average throwout by band. Each point represents the throwout rate for a single word; they are randomly jittered in the x axis only for easier display.

$\hat{y}_0$ threshold	$\mathbb{P}(Y < 0)$	$\mathbb{P}(Y \in [0, 0.5))$	$\mathbb{P}(Y > 1)$	$\mathbb{P}$(throwout of $Y > 1$)	# accepted
-0.755	0.151	0.312	0.190	0.000	67833
⋮					
0.495	0.091	0.291	0.228	0.110	50399
0.520	0.083	0.282	0.237	0.145	46755
0.545	0.077	0.274	0.247	0.184	42931
0.570	0.071	0.265	0.257	0.234	38758
0.595	0.065	0.254	0.269	0.288	34562
0.620	0.057	0.246	0.282	0.351	30208
0.645	0.053	0.232	0.297	0.418	25867
0.670	0.047	0.220	0.312	0.488	21735
0.695	0.044	0.207	0.329	0.561	17873
0.720	0.039	0.192	0.350	0.635	14160
0.745	0.033	0.175	0.370	0.704	10905
0.770	0.030	0.157	0.389	0.772	8039
0.795	0.025	0.142	0.414	0.831	5670
0.820	0.022	0.126	0.447	0.877	3855
0.845	0.019	0.118	0.474	0.916	2491
0.855	0.019	0.112	0.477	0.930	2048
0.865	0.018	0.103	0.491	0.941	1694
0.875	0.017	0.101	0.509	0.951	1381
0.885	0.013	0.096	0.524	0.959	1100
0.895	0.010	0.086	0.541	0.966	909
0.905	0.008	0.078	0.552	0.973	717
0.915	0.009	0.074	0.573	0.978	571
0.925	0.004	0.057	0.604	0.981	454
0.935	0.006	0.054	0.605	0.985	354
0.945	0.008	0.057	0.630	0.989	262
0.955	0.010	0.060	0.648	0.991	199

TABLE 2: Oos RF performance results for the [word unseen] system operating under a variety of $\hat{y}_0$ thresholds (oos $R^2 = 0.177$). We compute empirical cumulative distribution function values of interest as well as the cost (the throwout rate of informative contexts). We also display the number of contexts marked as acceptable of the 67,833 contexts in the original training data. At low thresholds (below 0.5), the system keeps all contexts, so these thresholds are not relevant and are colored gray. Past a threshold of 0.925, we have 98.1% throwout of our good contexts and acceptability is so low, we lose the ability to accurately estimate empirical probabilities; they are also colored gray. Cutoffs usable in practice are colored yellow (see main text). The last column displays the number of contexts accepted from the 67,833 in the training data to inform intuition on the confidence intervals of the estimated probabilities; low values are not stable.

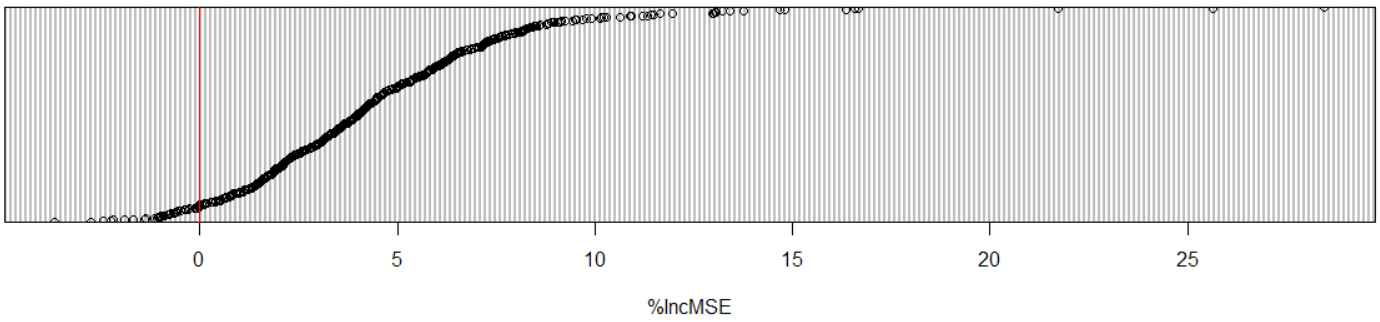


Fig. 6: Variable importance for all 615 features as measured by the increase in percent mean squared error. The red line is the cutoff for features that no longer increase out-of-bag accuracy of the RF model. 50 features (8%) are in this category. Feature names are omitted due to space restrictions.

between the features by permutating values of one feature but keeping the other feature values the same.

We learn here that the vast majority of features contribute synergistically to oos predictive performance. If the model had only a few variables contributing, there would be many more scores with near zero or negative oos mean squared error increase.

It is possible we could drop some of the 50 non-contributing variables in a stepwise fashion without performance loss, but running the stepwise elimination would be computationally prohibitive. Thus, Figure 6 cannot answer the question "how many variables matter?" without further work.

Which variables likely contribute the most to performance? In Figure 7, we illustrate the top 30 most important variables of Figure 6 (the right tail) and print their variable names. We now describe the top seven variables and speculate as to why they contain the most information about context quality.

1. `similar_1_10` tallies the number of the most similar word stems to the target word (the top 10 most similar as returned by the DISCO system querying the COCA

$\hat{y}_0$ threshold	$\mathbb{P}(Y<0)$	$\mathbb{P}(Y\in[0,0.5))$	$\mathbb{P}(Y>1)$	$\mathbb{P}$(throwout of $Y>1$)	# accepted
0.445	0.091	0.297	0.226	0.110	51116
0.470	0.086	0.293	0.231	0.136	48696
0.495	0.082	0.289	0.236	0.165	46111
0.520	0.078	0.283	0.242	0.196	43356
0.545	0.072	0.278	0.249	0.231	40424
0.570	0.068	0.272	0.256	0.272	37352
0.595	0.064	0.265	0.264	0.314	34259
0.620	0.062	0.257	0.273	0.357	31130
0.645	0.057	0.250	0.282	0.406	27992
0.670	0.054	0.243	0.290	0.457	24921
0.695	0.051	0.236	0.300	0.507	21964
0.720	0.048	0.230	0.310	0.557	19113
0.745	0.046	0.222	0.320	0.610	16367
0.770	0.043	0.216	0.332	0.657	13999
0.795	0.040	0.208	0.343	0.699	11899
0.820	0.039	0.200	0.355	0.741	9953
0.845	0.035	0.191	0.369	0.778	8227
0.870	0.032	0.183	0.383	0.811	6788
0.895	0.032	0.174	0.394	0.840	5614
0.920	0.029	0.168	0.410	0.868	4492
0.945	0.028	0.160	0.424	0.893	3563
0.970	0.027	0.155	0.438	0.913	2856
0.995	0.027	0.152	0.443	0.928	2307
1.020	0.026	0.145	0.454	0.942	1802
1.045	0.024	0.149	0.459	0.954	1411
1.070	0.025	0.149	0.474	0.963	1096
1.095	0.022	0.146	0.491	0.971	862
1.345	0.020	0.177	0.503	0.995	147
1.370	0.015	0.174	0.492	0.996	132
1.395	0.017	0.197	0.453	0.996	117
1.420	0.019	0.202	0.452	0.997	104
1.570	0.015	0.191	0.412	0.998	68
1.670	0.000	0.152	0.435	0.999	46

TABLE 3: Oos linear model performance results for the [word unseen] system operating under $\hat{y}_0$ thresholds which produce similar $\mathbb{P}(Y<0)$ quantiles as Table 2. Here, oos $R^2=0.167$ which is approximately the same as the RF model but performance based on our holistic cost metric is considerably lower. Again, R^2 is not an appropriate guage of fit for our model's performance.

$\hat{y}_0$ threshold	$\mathbb{P}(Y<0)$	$\mathbb{P}(Y\in[0,0.5))$	$\mathbb{P}(Y>1)$	$\mathbb{P}$(throwout of $Y>1$)	# accepted
0.840	0.019	0.122	0.466	0.881	2259
0.845	0.020	0.120	0.471	0.890	2068
0.855	0.020	0.113	0.474	0.908	1720
0.865	0.019	0.102	0.492	0.921	1434
0.880	0.014	0.095	0.521	0.940	1038
0.890	0.011	0.096	0.529	0.949	875
0.895	0.010	0.085	0.542	0.953	788

TABLE 4: Same as Table 2 but now we only look at oos results for bands 1–6. Note the proportion of directive contexts is about the same but throwout has improved substantially.

corpus). This information is likely important because the reiteration of words related to the meaning of the target word provides contextual clues.

2. `collocation_1_10` is the same as `similar_1_10` except it tallies the top most collocated words. Similar to words with related meaning, these words aid in scoping and limiting the meaning of the target word.
3. `politeness_component` is a feature returned by the Sentiment Analysis and Cognition Engine [65] that measures politeness using the dictionary lists of "polite" words found in Stone's [69] General Inquirer and Lasswell and Namenwirth's [70] dictionary lists. As to why this feature is important for informativeness, we do not know, but the explanation to number 7 below may apply.
4. `Kuperman.AoA.AW` is a feature returned by the Tool for the Automatic Analysis of Lexical Sophistication [61] that tallies the age of acquisition (a scale created by [40]) for all words in the context. This is likely a proxy for difficulty of the context.
5. `Kuperman.AoA.CW` is the same as above except it tallies for content words only. We assume these two features are collinear (future work can verify), thereby sharing places in the split rules in the Random Forest. If so, then Kuperman's scale should occupy the number 1 position.
6. `count.word1.target.word2` is a feature computed from our trigrams database illustrated on line 5 of Figure 2. Higher counts indicate the specific use in this context is prevalent. These two words sandwiching the target word likely provide information about its meaning.
7. `MRC.Meaningfulness.CW` is a feature returned by the Tool for the Automatic Analysis of Lexical Sophis-

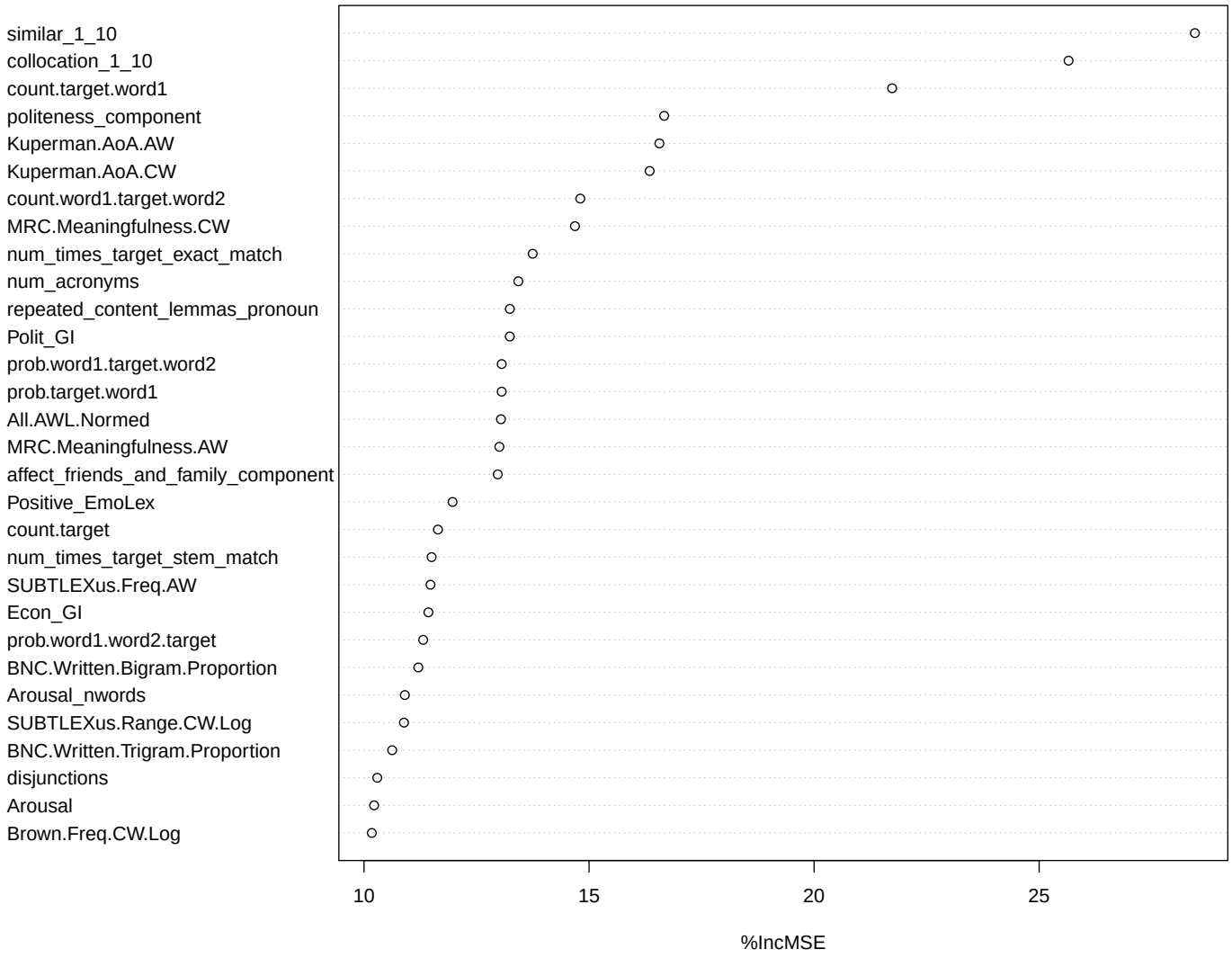


Fig. 7: Variable importance for the top 30 most important features as measured by the increase in percent mean squared error. Descriptions of each individual variable is found in Appendix C.

tication [61] that tallies the number of content words that are included among the Medical Research Council Psycholinguistic Database of English words [71] related to "meaningfulness". This score has been shown to be correlated to both writing quality [72] and lexical proficiency [73]. High values may explain why these contexts are rated highly by the Turkers, but it is our intuition that it does not truly relate to informativeness.

Higher likely corresponds to better to all of the above features except the Kuperman metrics. As a demonstration, consider the context below that was rated as highly informative ($y = 1.6$) for the band 8 word "infrastructure".

> Yet if we look closely at our transportation system, we see that the broad term "infrastructure" covers a dazzling variety of technologies serving very different needs. From cow paths to eight-lane expressways, from cars to trucks to barges to supersonic transports, our transportation infrastructure means many different technologies carrying many types of traffic at widely varying speeds. ... [74, Section 1]

Due to the presence of words "transportation", "transport", "expressways", "traffic", etc., the `similar_1_10` value was 5 (a z-score of +6.4) and the `collocation_1_10` value was 3 (a z-score of +3.8) meaning that this context is chock full of related words. The `politeness_component` scored 0.047 (a z-score of -0.8) i.e. about average. The `Kuperman.AoA.AW` value was 6.15 (a z-score of +0.4) and the `Kuperman.AoA.CW` value was 7.22 (a z-score of +0.2) indicating about average word difficulty and hence smooth readability: the reader won't be overburdened with additional new words especially in this context of a band 8 target word. The `count.word1.target.word2` value was 14221 (a z-score of -0.03) meaning that an average number texts shared the trigram "transportation infrastructure means" suggesting that this is not a nonstandard contextual usage. Finally, the `MRC.Meaningfulness.CW` metric was 470.6 (a z-score of +2.2) indicating this context's writer had above average proficiency and writing ability.

This exercise illustrates the main thrust of the technology — by boiling down a context into numeric features that

correlate with informativeness, contexts can be sorted based on educational value. Space limitations preclude us from discussing more features, the features in greater depth and how our RF model uses feature interactions. There is no doubt much can be learned about the "why" of context informativeness by querying our model.

3.5 Ordering of the Contexts

As discussed during our problem setup, we prefer uninformative contexts to appear at the later stages of learning. To test this, we use the model in Table 2 corresponding to the threshold of 0.845. We examine all words with five or more contexts in the oos use distribution and plot the probability of seeing an uninformative or misdirective context by order of appearance once we order by predicted informativeness best to worst (Figure 8). As shown, about 47% of the words ever had such a context and it was uncommon to see that type of context early on; only 11% of words in the first two exposures and 31% within the first 5 exposures.

4 External Teacher Validation

The stated purpose of our system is to automatically identify informative contextual examples for vocabulary instruction in high school students. To externally validate the quality of our [word unseen] system's output, we conducted a randomized experiment with high school teachers.

4.1 Methods

Participants included 31 high school language arts teachers from the United States (30 from South Carolina and 1 from Connecticut). They were recruited through social media advertising and an email campaign. Each participant was asked to complete a web-based survey asking basic demographic information and 18 experimental questions.

Each participant was shown different experimental questions created as follows. First, three target words were randomly selected without replacement from each of the first six difficulty bands for a total of 18 unique words. For each word, one context was drawn at random from the original DictionarySquared database (uncurated) and one context was drawn at random from the future use distribution, the set predicted to be used by [word unseen] model at $\hat{y} = 0.895$ (curated). Put another way, one context was drawn randomly from the red distribution in Figure 4 and one is drawn randomly from the green distribution. For each word, the teacher was asked which context would be better for teaching the word. The two contexts were presented randomly side-by-side below the prompt. A screenshot of the experimental question is provided in Figure 9. Each question was a separate web page and teachers were not allowed to change responses upon each page submission.

Collecting many comparisons of a curated context and an uncurated context drawn at random is an honest way to test if our out-of-sample results of Table 2 comport with professional language arts educators' preferences. Here we are testing the superiority of the median informativeness of the set of contexts sifted by our model.

27 participants answered all 18 questions, and the remaining 4 participants answered between 1 and 11 questions for a total of $n = 502$ trials. The 27 teachers that completed the survey were rewarded with a $5 Amazon gift card, an incentive that was known before their participation.

4.2 Results

Of the 502 trials, the curated set was selected 305 times or 60.7%, a significant result when testing the null hypothesis that there is no preference between the curated and uncurated contexts ($p \approx 1.8 \times 10^{-6}$). This test is only valid if all teachers' responses are statistically independent. To test this assumption, we tested the differential performance between the teachers using a logistic regression with dummy variables for each teacher. Pitting this model versus a simple model of only an intercept turns up insignificant via the likelihood ratio test ($p = 0.46$). Thus, there is no reason to believe the assumption of independence is not justified.

These results provide unequivocal external validation that our [word unseen] model selects contexts that are more suitable for teaching than the original DictionarySquared database. However, it may seem that the teachers' choices of the curated set 60.7% of the time is low — you may wonder why the teachers could not select the context from the curated set 100% of the time. An analysis of Figure 4 will demonstrate that perfect discrimination is not possible: the average context in the uncurated set is between nondirective and general and due to random sampling, the context from the uncurated set can be more informative than the context from the curated set. When running simulations, 60.7% is in the ballpark of expected discrimination, especially when considering the fact that there are inevitable judgment calls when the two contexts are similar in informativeness.

5 Discussion

Considering our RF model for the [word unseen] system, we argue that our predictive performance is good enough to implement this system in a context-collection effort without the need for a human rater. However, we limit our unconditional recommendation to future target words within the level defined by our band 1–6, words that were externally validated by high school teachers and words of which the throwout rate is not overly punitive.

The following example may demonstrate a typical final product of our model's curation. For the target word "malevolent" in band 6, the following context

> From Scotland comes stories of the Old Hag or Night Hag. The Old Hag is a **malevolent** spirit that visits people in the middle of the night while they sleep. Those who survive this nocturnal visit report being awoken with a feeling of dread or unease but unable to move or speak. [75, third paragraph]

received an average MTurk rating of 0.7 i.e. directive and highly informative as we can see from all the contextual clues. The RF oos prediction here is 0.89. Let's compare this to

> If you or someone you know has gotten nothing but heartache this Valentine's Day (or any other occasion involving that **malevolent** blood-pumping

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2018.2789900, IEEE Transactions on Learning Technologies

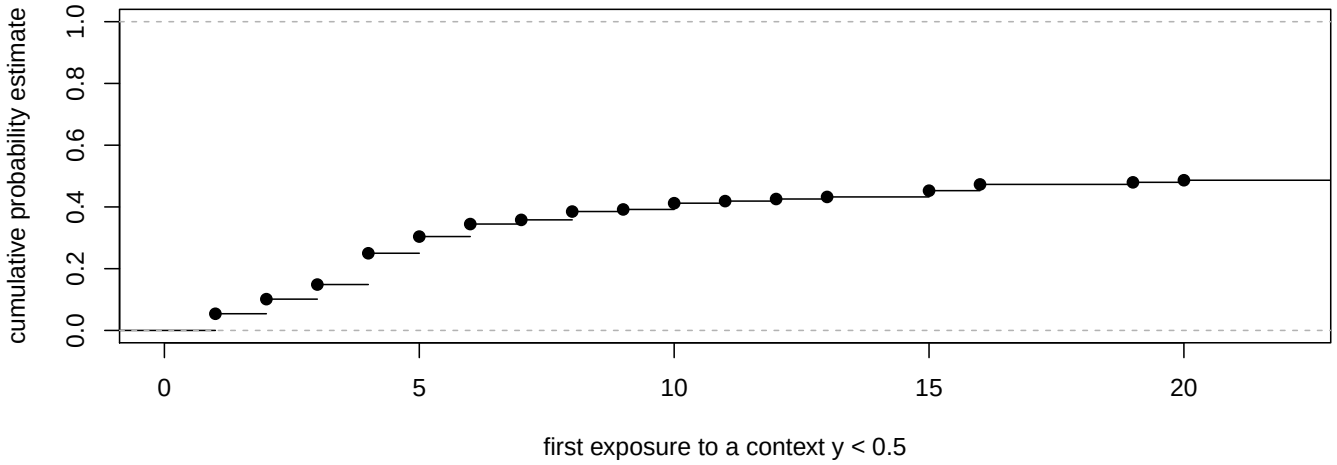


Fig. 8: Under the use distribution, the cumulative probability distribution of the first exposure to a context with a true rating less than 0.5 (non-informative or misdirective). We limit words here to those with 5 or more contexts.

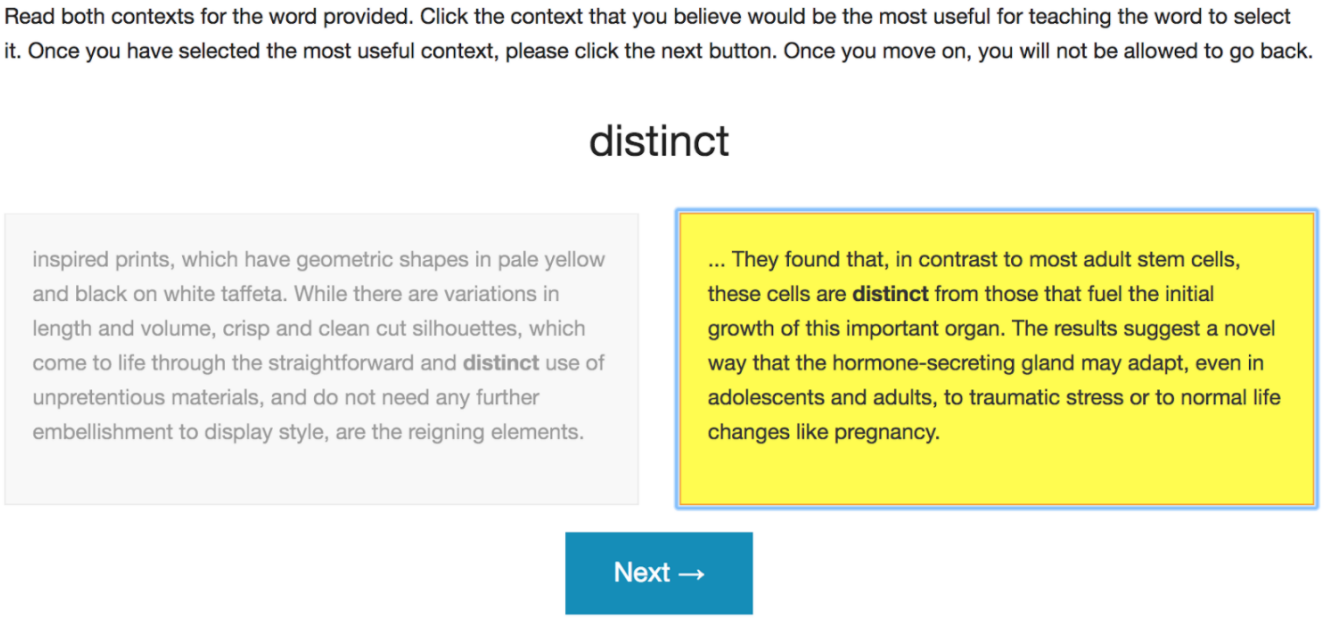


Fig. 9: A screenshot of a typical experimental question for the teacher validation study. Here, the target word was "distinct" and the teacher has already selected the context on the right by clicking on its box. By pressing "Next →", the choice would be finalized and the next question would be presented.

> organ), this 12-song collection offers the perfect antidote. Includes the J. Geils Band's immortal "Love Stinks," Gram Parsons' defining version of "Love Hurts," Joy Division's "Love Will Tear Us Apart," and more. [76, fourth paragraph]

which received an average MTurk rating of -0.1 indicating it is non-directive and possibly even misdirective. Here, the RF oos prediction is 0.48.

Thus, when implementing the RF model for the [word unseen] system in Table 4 for a highlighted threshold, the first context would be administered to students in our vocabulary instruction system and the second would not be administered.

The last point to discuss is the high throwout rates of our best contextual examples of the target word. In order for this system to be practical, we would query massive corpora (such as the Internet) for contextual examples and we would optimize our routines that compute the 615 features. Both are possible and thus we do not anticipate the high throwout rate to be a problem in practice.

Once again, we have developed a system that has the ability to automatically identify informative contexts for learning arbitrary words of interest and our technology can be greatly beneficial to educators and researchers.

5.1 Future Directions

This work represents our initial steps toward automatic identification of useful contexts for vocabulary learning.

Here we note three areas for future work.

First, we would like to extend the model by considering other relevant features. Given that nearly all of our already-tested features were found to be important in the RF models, we are nearly certain that other intelligent features can provide additional predictive power due to our current $R^2 \approx 20\%$.

Second, although the current system is useful, it is quite cumbersome to obtain a large corpus, calculate 615 features, and subsequently discard 85–95% of retrieved contexts in order to identify useful ones. As we continue to explore new features that can be added, we like to give further consideration to the costs and benefits of models that use fewer features, perhaps ordered by ease of feature calculation. Reducing the effort required to obtain predictions of context informativeness would likely increase the practical utility of this approach.

Third, selecting features believed to be predictive of the response is known as "hand-engineering" and claimed to be a failing due to arbitrariness of the specific features collected as well as non-generalizability of the specific features to other tasks [77]. However, given the relative uncharted territory of predicting informativeness (in comparison to the well-trod NLP problems of "part-of-speech tagging" or "named entity recognition"), we believe this to be a good first pass at the problem that we hope will be iteratively improved. State-of-the-art systems for solving NLP problems seem to be gravitating towards deep learning [77], using many-layered neural networks operating on the raw text data itself as well as clever text representations (see e.g. the work of Socher [78]). This may help us employ features without the need for explicitly specifying them, a concept known as *representation learning*. These learned representations often result in much better performance than can be obtained with our strategy herein of hand-designed features [56]. Such strategies are left to future work and may even be synthesized together with ours in an ensemble "superlearner" [79].

Data

The full training data set and the teacher validation dataset can be found at github.com/kapelner/predicting_contextual_informativeness with the GPL3 license.

Acknowledgments

We would like to thank Julie Byard for help with eliciting the MTurk ratings, Jack Mostow for helpful discussions and code from his work, Charles Perfetti, Margaret McKeown and Joanna Scoggins for careful reads of the manuscript. Additionally, Abba Krieger, Justin Bleich and Richard Berk provided invaluable advice. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130467 to the University of South Carolina (Adlof, PI). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

[1] D. Braze, W. Tabor, D. P. Shankweiler, and W. E. Mencl, "Speaking up for vocabulary reading skill differences in young adults," *Journal of Learning Disabilities*, vol. 40, no. 3, pp. 226–243, 2007.

[2] J. B. Carroll, *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press, 1993.

[3] J. Lee, "Size matters: Early vocabulary as a predictor of language and literacy competence," *Applied Psycholinguistics*, vol. 32, no. 01, pp. 69–92, 2011.

[4] S. A. Storch and G. J. Whitehurst, "Oral language and code-related precursors to reading: evidence from a longitudinal structural model." *Developmental Psychology*, vol. 38, no. 6, p. 934, 2002.

[5] C. A. Perfetti, M. A. Britt, and M. C. Georgi, *Text-based learning and reasoning: Studies in history*. Routledge, 2012.

[6] M. F. Hock, I. F. Brasseur, D. D. Deshler, H. W. Catts, J. G. Marquis, C. A. Mark, and J. W. Stribling, "What is the reading component skill profile of adolescent struggling readers in urban schools?" *Learning Disability Quarterly*, vol. 32, no. 1, pp. 21–38, 2009.

[7] S. M. Brusnighan and J. R. Folk, "Combining contextual and morphemic cues is beneficial during incidental vocabulary acquisition: Semantic transparency in novel compound word processing," *Reading Research Quarterly*, vol. 47, no. 2, pp. 172–190, 2012.

[8] S. M. Brusnighan, R. K. Morris, J. R. Folk, and R. Lowell, "The role of phonology in incidental vocabulary acquisition during silent reading," *Journal of Cognitive Psychology*, vol. 26, no. 8, pp. 871–892, 2014.

[9] K. Cain, J. Oakhill, and K. Lemmon, "Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity." *Journal of educational psychology*, vol. 96, no. 4, p. 671, 2004.

[10] R. Chaffin, R. K. Morris, and R. E. Seely, "Learning new word meanings from context: A study of eye movements," *Journal of Experimental Psychology Learning Memory and Cognition*, vol. 27, no. 1, pp. 225–235, 2001.

[11] G. A. Frishkoff, K. Collins-Thompson, C. A. Perfetti, and J. Callan, "Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment," *Behavior Research Methods*, vol. 40, no. 4, pp. 907–925, 2008.

[12] R. K. Morris and R. S. Williams, "Bridging the gap between old and new: Eye movements and vocabulary acquisition in reading," in *The mind's eye: Cognitive and applied aspects of eye movement research*, J. Hyona, Ed. Amsterdam: Elsevier Science BV, 2003, ch. 12, pp. 235–252.

[13] W. E. Nagy, R. C. Anderson, and P. A. Herman, "Learning word meanings from context during normal reading," *American Educational Research Journal*, vol. 24, no. 2, p. 237, 1987.

[14] R. Williams and R. Morris, "Eye movements, word familiarity, and vocabulary acquisition," *European Journal of Cognitive Psychology*, vol. 16, no. 1-2, pp. 312–339, 2004.

[15] B. Laufer, "Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? some empirical evidence," *Canadian modern language review*, vol. 59, no. 4, pp. 567–587, 2003.

[16] W. E. Nagy and R. C. Anderson, "How many words are there in printed school english?" *Reading Research Quarterly*, vol. 19, no. 3, p. 304, 1984.

[17] A. E. Cunningham and K. E. Stanovich, "Early reading acquisition and its relation to reading experience and ability 10 years later." *Developmental Psychology*, vol. 33, no. 6, p. 934, 1997.

[18] S. E. Mol and A. G. Bus, "To read or not to read: a meta-analysis of print exposure from infancy to early adulthood." *Psychological Bulletin*, vol. 137, no. 2, p. 267, 2011.

[19] K. E. Stanovich, R. F. West, and M. R. Harrison, "Knowledge growth and maintenance across the life span: The role of print exposure." *Developmental Psychology*, vol. 31, no. 5, p. 811, 1995.

[20] I. L. Beck, M. G. McKeown, and L. Kucan, *Bringing words to life: Robust vocabulary instruction*. Guilford Press, 2013.

[21] J. F. Baumann, E. J. Kame'enui, and G. E. Ash, "Research on vocabulary instruction: Voltaire redux," in *Handbook of research on teaching the English language arts (2nd edition)*, J. Flood, D. Lapp, J. R. Squire, and J. Jensen, Eds. Mahwah, NJ: Lawrence Erlbaum, 2003, pp. 752–785.

[22] S. A. Stahl and M. M. Fairbanks, "The effects of vocabulary instruction: A model-based meta-analysis," *Review of Educational Research*, vol. 56, no. 1, p. 72, 1986.

[23] E. D. Reichle and C. A. Perfetti, "Morphology in word identification: A word-experience model that accounts for morpheme frequency effects," *Scientific Studies of Reading*, vol. 7, no. 3, pp. 219–237, 2003.

[24] W. E. Nagy and J. A. Scott, "Vocabulary processes," in *Handbook of Reading Research (Volume 3)*, M. L. Kamil, P. Mosethal, P. D. Pearson, and R. Barr, Eds. Mahway, NJ: Erlbaum, 2000, pp. 269–284.

[25] S. Adlof, M. McKeown, C. Perfetti, A. Kapelner, S. Nessaiver, and J. Soterwood, "Dictionarysquared development paper," 2016, working paper.

[26] I. L. Beck, M. G. McKeown, and E. S. McCaslin, "Vocabulary development: All contexts are not created equal," *The Elementary School Journal*, vol. 83, no. 3, p. 177, jan 1983.

[27] G. A. Frishkoff, C. Perfetti, and K. Collins-Thompson, "Predicting robust vocabulary growth from measures of incremental learning," *Scientific Studies of Reading*, vol. 15, no. 1, pp. 71–91, jan 2011.

[28] S. Adlof, G. Frishkoff, J. Dandy, and C. Perfetti, "Effects of induced orthographic and semantic knowledge on subsequent learning: a test of the partial knowledge hypothesis," *Reading and Writing*, vol. 29, no. 3, pp. 475–500, 2016.

[29] E. Silverstein. (n.d.) Avoiding seasickness. [Online]. Available: http://www.cruisecritic.com/articles.cfm?ID=48

[30] M. Ingram. (n.d.). [Online]. Available: http://www.woebot.com/movabletype/archives/000138.html

[31] J. Brown and M. Eskenazi, "Retrieval of authentic documents for reader-specific lexical practice," in *InSTIL/ICALL Symposium 2004*, 2004.

[32] K. Collins-Thompson and J. Callan, "A language modeling approach to predicting reading difficulty." in *Proceedings of NAACL HLT*, 2004.

[33] J. Mostow and W. Duan, "Generating example contexts to illustrate a target word sense," in *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2011, pp. 105–110.

[34] S. Hassan and R. Mihalcea, "Learning to identify educational materials," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 8, no. 2, p. 2, 2011.

[35] S. S. Nash, "Learning objects, learning object repositories, and learning theory: Preliminary best practices for online courses," *Interdisciplinary Journal of Knowledge and Learning Objects*, vol. 1, pp. 217–228, 2005.

[36] J. Mostow, D. Gates, R. Ellison, and R. Goutam, "Automatic identification of nutritious contexts for learning vocabulary words," *International Educational Data Mining Society*, 2015.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, tenth printing, second ed. Springer, 2009.

[38] A. Coxhead, "A new academic word list," *TESOL Quarterly*, vol. 34, no. 2, pp. 213–238, 2000.

[39] S. Zeno, S. Ivens, R. Millard, and R. Duvvuri, *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates, 1995.

[40] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert, "Age-of-acquisition ratings for 30,000 english words," *Behavior Research Methods*, vol. 44, no. 4, pp. 978–990, 2012.

[41] I. L. Beck, M. G. McKeown, and R. C. Omanson, "The effects and uses of diverse vocabulary instructional techniques," in *The nature of vocabulary acquisition*, M. McKeown and M. E. Curtis, Eds. Hillsdale, NJ: Erlbaum, 1987, pp. 147–163.

[42] P. Nation, "Learning vocabulary in lexical sets: Dangers and guidelines," *TESOL Journal*, vol. 9, no. 2, pp. 6–10, 2000.

[43] S. A. Stahl and W. E. Nagy, *Teaching word meanings*. Routledge, 2007.

[44] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 254–263.

[45] E. Gibson, S. Piantadosi, and K. Fedorenko, "Using mechanical turk to obtain and analyze english acceptability judgments," *Language and Linguistics Compass*, vol. 5, no. 8, pp. 509–524, 2011.

[46] J. K. Goodman, C. E. Cryder, and A. Cheema, "Data collection in a flat world: The strengths and weaknesses of mechanical turk samples," *Journal of Behavioral Decision Making*, vol. 26, no. 3, pp. 213–224, 2013.

[47] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on amazon mechanical turk," *Judgment and Decision making*, vol. 5, no. 5, pp. 411–419, 2010.

[48] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2008, pp. 453–456.

[49] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, 2010, pp. 64–67.

[50] D. M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 867–872, 2009.

[51] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 491–518, 2012.

[52] T. M. Byun, P. F. Halpin, and D. Szeredi, "Online crowdsourcing for efficient rating of speech: A validation study," *Journal of Communication Disorders*, vol. 53, pp. 70–83, 2015.

[53] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[54] M. Bilenko, A. Kamenev, V. Narayanan, and P. Taraba, "Adaptive featurization as a service," January 2016, uS Patent 20,160,012,318.

[55] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[56] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[57] A. Franz and T. Brants. (2006) Official google research blog: All our n-gram are belong to you. [Online]. Available: http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

[58] P. Kolb, "DISCO: A multilingual database of distributionally similar words," *Proceedings of KONVENS-2008, Berlin*, 2008.

[59] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1995, pp. 189–196.

[60] M. Davies, "The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights," *International Journal of Corpus Linguistics*, vol. 14, no. 2, pp. 159–190, 2009.

[61] K. Kyle and S. A. Crossley, "Automatically assessing lexical sophistication: Indices, tools, findings, and application," *TESOL Quarterly*, vol. 49, no. 4, pp. 757–786, 2015.

[62] B. Laufer, "What's in a word that makes it hard or easy? intralexical factors affecting the difficulty of vocabulary acquisition," in *Vocabulary Description, Acquisition and Pedagogy*, M. McCarthy and N. Schmitt, Eds. Cambridge, United Kingdom: Cambridge University Press, 1997, ch. 2.3, pp. 140–155.

[63] S. A. Crossley, K. Kyle, and D. S. McNamara, "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion," *Behavior Research Methods*, pp. 1–11, 2015.

[64] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, 2014.

[65] S. A. Crossley, K. Kyle, and D. S. McNamara, "Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis," *Behavior Research Methods*, pp. 1–19, 2016.

[66] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998, pp. 92–100.

[67] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[68] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., 1997.

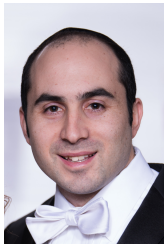
[69] P. J. Stone, D. C. Dunphy, and M. S. Smith, *The general inquirer: A computer approach to content analysis.* MIT press, 1966.

[70] H. D. Lasswell and J. Z. Namenwirth, *The Lasswell value dictionary*. New Haven: Yale University Press, 1969.

[71] M. Coltheart, "The MRC psycholinguistic database," *The Quarterly Journal of Experimental Psychology*, vol. 33, no. 4, pp. 497–505, 1981.

[72] S. A. Crossley and D. S. McNamara, "Understanding expert ratings of essay quality: Coh-metrix analyses of first and second language writing," *International Journal of Continuing Engineering Education and Life Long Learning*, vol. 21, no. 2-3, pp. 170–191, 2011.

[73] S. A. Crossley, T. Salsbury, and D. S. McNamara, "Predicting the proficiency level of language learners using lexical indices," *Language Testing*, vol. 29, no. 2, pp. 243–263, 2012.
[74] M. A. Sirbu. (n.d.) Telecommunications technology and infrastructure. [Online]. Available: https://www.cs.cmu.edu/afs/andrew/usr9/sirbu/www/pubs/ipaper.html
[75] "Terror in the Night". (n.d.). [Online]. Available: http://www.hauntedbay.com/features/nightterror.shtml
[76] S. Elliott. (2006) Valetine's day. [Online]. Available: http://newappeal.blogspot.com/2006/02/valetines-day.html
[77] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*. ACM, 2008, pp. 160–167.
[78] R. Socher, "Recursive deep learning for natural language processing and computer vision," Ph.D. dissertation, Citeseer, 2014.
[79] E. C. Polley and M. J. van der Laan, "Super learner in prediction," U.C. Berkeley Division of Biostatistics Working Paper Series, Berkeley, CA, Tech. Rep. Paper 266, 2010.



Adam Kapelner Adam Kapelner received his Ph.D. in Statistics from the Wharton School of the University of Pennsylvania. He is an Assistant Professor of Mathematics at Queens College, City University of New York. His research interests include machine learning — their algorithms and prediction applications, statistical methodology for experimentation, crowdsourcing and more generally, engineering systems using Statistics to solve various applied problems.

Jeanine Soterwood Jeanine Soterwood received her Ph.D. in Applied Mathematics from the University of Arizona. She is the principal consultant at Littleforest Consulting, a software development consulting firm. Her research interests include data science with an emphasis on machine learning applications and data derived from natural language processing methods.

Shalev Nessaiver Shalev NessAiver is a freelance programmer. His research interests include pedagogy, education technology, cognition, and human computer interaction.

Suzanne Adlof Suzanne Adlof received her Ph.D. in Speech Language Pathology from the University of South Carolina. She is an Assistant Professor of Communication Sciences and Disorders at the University of South Carolina. Her research interests include understanding the relationship between oral written language development, identifying reading and language impairments, and developing effective interventions to improve reading and academic outcomes, and she has published numerous articles and chapters on these topics. She is a voting member of the Society for the Scientific Study of Reading as well as the American Speech-Language-Hearing Association.