Can Textbook Annotations Serve as an Early Predictor of Student Learning?

Adam Winchell
University of Colorado Boulder
adam.winchell@colorado.edu

Michael Mozer
University of Colorado Boulder
mozer@colorado.edu

Andrew Lan
Princeton University
andrew.lan@princeton.edu

Phillip Grimaldi
OpenStax Foundation
phillip.grimaldi@rice.edu

Harold Pashler
UCSD
hpashler@ucsd.edu

ABSTRACT

When engaging with a textbook, students are inclined to highlight key content. Although students believe that highlighting and subsequent review of the highlights will further their educational goals, the psychological literature provides no evidence of benefits. Nonetheless, a student's choice of text for highlighting may serve as a window into their mental state—their level of comprehension, grasp of the key ideas, reading goals, etc. We explore this hypothesis via an experiment in which 198 participants read sections from a college-level biology text, briefly reviewed the text, and then took a quiz on the material. During initial reading, participants were able to highlight words, phrases, and sentences, and these highlights were displayed along with the complete text during the subsequent review. Consistent with past research, the amount of highlighted material is unrelated to quiz performance. However, our main goal is to examine highlighting as a data source for inferring student understanding. We explored multiple representations of the highlighting patterns and tested Bayesian linear regression and neural network models, but we found little or no relationship between a student's highlights and quiz performance. Our long-term goal is to design digital textbooks that serve not only as conduits of information into the mind of the reader, but also allow us to draw inferences about the reader at a point where interventions may increase the effectiveness of the material.

Keywords

student modeling, bayesian regression, neural networks

1. INTRODUCTION

A premise of educational data mining is that the knowledge state of a student can be inferred by observation. However, knowledge state is opaque until students reach a level of understanding that they can be tested or they can solve problems. This delay makes interventions at an early stage of exposure quite challenging. Consider a student's first engagement with new material in a textbook. Reading times and gaze patterns may be useful for modeling student engagement and comprehension [3]. However, these implicit measures are quite difficult to collect. Fortunately, students often willingly provide explicit measures: students will voluntarily highlight sections of text and write notes in the margins. With the advent of electronic texts, the opportunity now exists to collect data from students during their early exposure to new material, and if knowledge state can be inferred, interventions can be performed early. In this article, we explore the hypothesis that these annotations—specifically highlights—can be used to predict comprehension, as assessed by a follow-up quiz.

Highlighting has been studied in the psychological literature from the perspective of whether highlighting is an effective study strategy. The current understanding is that the mere act of highlighting does not promote learning, nor does re-reading isolated sentences that were highlighted [1]. No relationship has been found between coarse statistics of highlighting (e.g. the total amount of text highlighted) and a student's performance/understanding [2].

In a few cases, highlighting has been shown to provide benefits. First, text which is pre-highlighted by an informed instructor can guide a student to focus on key content [4]. Second, restricting highlighting to encourage consideration of the material—e.g., permitting the student to highlight only one sentence per paragraph—can support understanding [5]. In contrast to this traditional literature that examines highlighting as a study tool, here we examine highlighting as a data source for inferring student understanding.

2. EXPERIMENT

We conducted an experiment in which participants read passages from a biology textbook. They later reviewed the passages, and then took a short quiz drawing on material from the passages. During initial reading, participants were allowed to highlight portions of the text (words, phrases, or sentences). These highlights were displayed along with the text during the review phase, and participants were instructed that highlighting could assist in the review.

2.1 Methodology

2.1.1 Participants

Participants aged 18 and above were recruited from Amazon Mechanical Turk. A total of 198 people completed the experiment and were paid $3.60. Data from six participants was discarded because these participants reported that they were unable to use the highlighting functionality in their web browser. The experiment took 25-30 minutes to complete. No screening was performed to determine an individual's background in biology. To incentivize attention to the task, participants were told that they would be entered into a raffle for a bonus prize of $15.00, with the number of entries equal to the number of correct reponses to the quiz questions.

2.1.2 Materials

Three passages were selected from the Openstax *Biology* textbook [7]. The passages were chosen with the expectation that they could be understood by a college-aged reader with no background in biology. The three passages concern the topic of sterilization, with one serving as an introduction, one discussing procedures, and the last summarizing commercial use. Twelve factual quiz questions were generated by selecting particular sentences from the passages and turning the factual statements in these sentences into fill-in-the-blank questions. These twelve questions were transformed into twelve additional multiple choice questions, each question comprised of the correct response and three lures as alternatives. Three questions are drawn from the first passage, four from the second passage, and five from the final passage.

For each participant a *normalized quiz score* is computed as follows. For each of the twelve questions, a score of 1.0 is assigned if both the fill-in-the-blank and multiple-choice response are correct; a score of 0.66 is assigned if the fill-in-the-blank (FIB) response is correct; a score of 0.33 is assigned if the multiple-choice (MC) response is correct; and a score of 0 is assigned if neither is correct. The normalized quiz score is the sum of these scores divided by 12, yielding a value in the range [0,1]. A liberal criterion was used for judging FIB response correctness: A response is considered correct if the edit distance between the actual and correct responses is less than 25% of the length of the correct response. Table 1 shows the distribution of response correctness on MC and FIB versions of a question.

2.1.3 Procedure

The experiment is divided into three phases. During the *reading* phase, the three passages are presented on the screen sequentially, each on screen for five minutes. During the *review* phase, the three passages are again presented sequentially, along along with any highlights the participant made during the reading phase, each for one minute. Finally, during the *quiz* phase, the 12 FIB questions are shown, followed by the 12 MC questions, randomized within question type. During the first two phases, a timer at the top of the screen indicates time remaining for the current passage. After the timer has expired, the screen blanks and displays a message describing the next step of the experiment (either the next passage or the next phase of the experiment). Throughout the course of the experiment, a progress bar is displayed at the bottom of the screen that indicates the current proportion of the experiment completed.

In the reading phase, participants may highlight text by selecting one or more words using the mouse, which we will refer to as a highlighting *interaction*. If the selected text exactly corresponds to an existing highlight, the highlight is deleted. If the selected text captures any portion of an existing highlight but extends beyond it, the existing highlight is expanded to include the new selection. A single interaction may highlight more than one sentence at a time, but cannot cross paragraph boundaries. In the review phase, the previously selected highlights are displayed, but no additional highlights can be made.

Table 1: Distribution of response correctness on multiple choice (MC) and fill-in-the-blank (FIB) versions of a question

	MC Incorrect	MC Correct
FIB Incorrect	0.259	0.415
FIB Correct	0.038	0.288

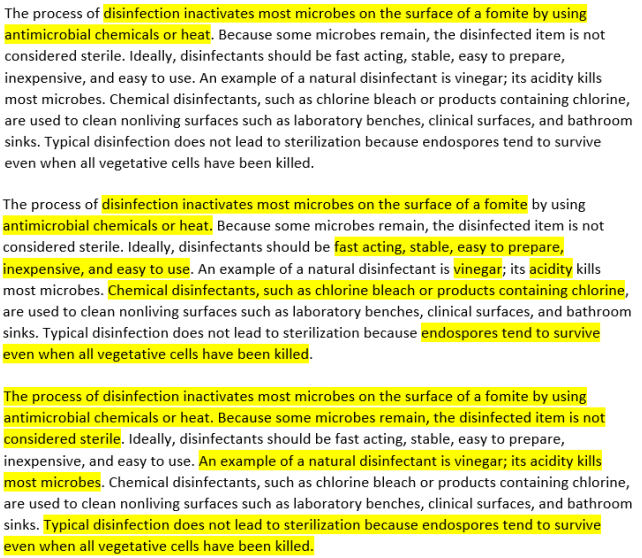

Figure 1: A paragraph of text as highlighted by three randomly selected participants.

3. RESULTS

Figure 1 presents an example of three participants' highlights of one paragraph of text. As these examples make clear, there is diversity in the manner in which individuals highlight. Highlights are used to note single words, phrases, and complete sentences.

In order to analyze the relationship between an individual's highlights and quiz performance, we need to first specify a representation of the highlights. In all analyses, we ignore the time course and sequence of actions that the participant took to create and/or delete highlights, and instead consider only the terminal highlighted state of each passage. The three passages contain 117 complete sentences delin-

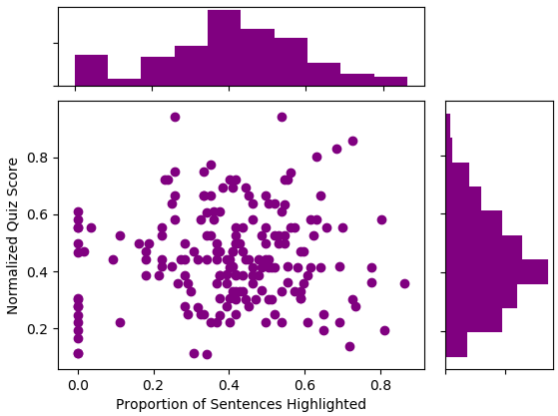

Figure 2: Scatter plot of proportion of sentences highlighted (using the binary encoding) versus normalized quiz score for each participant. The marginal distributions are shown above and to the right of the scatter plot.

eated by periods, exclamation marks, and question marks. The first analyses we perform are based on a *sentence-level* representation in which the pattern of highlights are coded as a 117-dimensional feature vector, either as a *binary* encoding in which each element i of the vector is set to 1 if any portion of sentence i is highlighted, or as a *graded* encoding in which element i is set to the proportion of words in the sentence that are highlighted.

Figure 2 shows the relationship between the proportion of sentences highlighted according to the binary encoding and the normalized quiz score. Each point is a single participant. As shown along the margin, the proportion of sentences highlighted is a unimodal distribution with a mean of 0.40. The normalized quiz score is also unimodal with a mean of 0.45. The scatter plot suggests no functional relationship—linear or otherwise—between the amount of highlighting and quiz performance; the correlation coefficient is 0.08.

Although the total number of highlights fails as a predictor of quiz score, the specific pattern of highlighting may prove more useful. To begin analyzing the relationship between highlighting patterns and performance, we performed a locally-linear embedding (LLE) with 11 neighbors [6] to reduce the dimensionality of the 117-dimensional binary sentence-level highlighting vector to a 2D space. Figure 3(a) plots the embedded points, colored to indicate the corresponding quiz score. The embedding has interesting structure, but no simple relationship to quiz performance. To understand what the LLE has captured, the points are recolored by proportion of sentences highlighted in Figure 3(b). This figure reveals that the abscissa captures the proportion, and the ordinate captures some of the diversity in the representation for a particular proportion. Referring back to Figure 3(a), there is no discernible relationship between the variation along the ordinate and performance, even when there is diversity in the embedding (i.e., the mid-range along the abscissa).

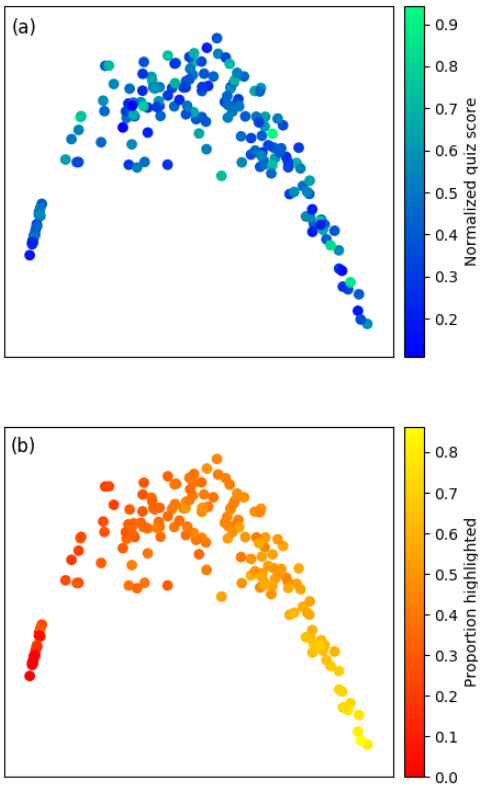

Figure 3: 2D LLE embedding of the binary sentence-level highlights with each point corresponding to one participant's data, and the coloring of points indicating (a) normalized quiz score and (b) the proportion of sentences highlighted.

We explored other parameterizations of LLE and other dimensionality reduction methods (e.g., k means clustering) but found no discernible relationship between performance and the reduced representations.

3.1 Modeling results

We constructed a series of models that map the highlighted-sentence representation—either the binary or graded encoding—to either total quiz score or correctness on specific problems. In all model testing, we perform nested cross validation to optimize model hyperparameters and evaluate model generalization to new participants. Our nested procedure consists of an outer 10-fold cross validation loop to partition the entire data set by participants into training and test sets, and an inner 3-fold cross validation loop further splitting the training set to select hyperparameters. The best set of hyperparameters chosen from the inner loop are selected and the entire training set is then used to build a model which predicts test set performance. This process is repeated over the outer loop to obtain an average normalized model loss.

The normalized model loss is defined as:

$$L = \frac{\mathrm{E}_i[\mathrm{E}_j[(s_{ij} - \hat{s}_{ij})^2]]}{\mathrm{E}_i[\mathrm{E}_j[(s_{ij} - \bar{s}_i)^2]]},$$

where s_{ij} is the test score of participant j in outer fold i, $\hat{s}_{ij}$ is the corresponding model prediction, $\bar{s}_i$ is the mean score of the participants in the training set for fold i, and E_i and E_j are the expectations taken over folds and participants, respectively. This normalized loss is 1.0 if the model does no better than predicting the mean score of the training participants, and drops to 0.0 if predictions are perfect. The proportion of variance explained by the model is $1 - L$.

3.1.1 Linear models

The first set of models we examine are based on Bayesian linear regression. The general form of these models is $y = Wx+b$, where y is the predicted quiz score, x is the highlight vector, and W and b are free parameters. This variant of linear regression is well suited for domains in which the number of input features is high relative to the volume of data. The model coefficients are regularized via a prior which achieves a ridge penalty (a penalty for large weights). The model is specified with a prior distribution on the precision of the observation noise, Gamma($shape = 10^{-\kappa_1}, rate = 10^{-\kappa_2}$) and prior on the precision of the coefficients, Gamma($shape = 10^{-\kappa_3}, rate = 10^{-\kappa_4}$), where $\kappa_* \in \{3, 4, 5, 6, 7, 8, 9\}$ was chosen by the inner cross validation loop.

Consistent with the scatter plot in Figure 2, when predicting on overall quiz performance, we found that a regression on the total number of sentence-level highlights obtains a mean normalized loss of 1.01 (SEM 0.0029). Similarly, a regression on the total number of words highlighted obtains a mean normalized loss of 1.01 (SEM 0.0028). Table 2 summarizes these and subsequent results.

Although the summary statistics fail to predict performance, one might hope to see a relationship between the specific *pattern* of highlights and performance. Unfortunately, regressions on the binary and graded sentence-level representations of highlights obtain mean normalized losses of 0.99 (SEM 0.0028) and 1.03 (SEM 0.0032), respectively.

We hypothesized that the sentence-level representations of highlights may be too coarse to capture important differences in the highlighting patterns. We therefore parsed the text based on a *sentence-fragment* representation in which the passages are segmented by periods, exclamation marks, question marks, colons, semicolons, as well as phrase-separating commas. The inclusion or exclusion of commas as segment boundaries was subjective; our strategy was to exclude commas that were used to delineate lists of items. Figure 4 gives an example of the sentence-fragment partition scheme. This partition scheme yields 235 sentence fragments across the three passages. We examined both binary and graded representations of the fragment-level highlights. The regression on the binary and graded fragment-level representations yields mean normalized losses of 0.93 (SEM 0.0024) and 1.03 (SEM 0.0032), respectively.

Parsing the passages at an even finer granularity, we constructed a representation of the individual *words* highlighted. The three passages have a total of 2291 word tokens. A regression on this raw representation of highlights yields a mean normalized loss of 0.93 (SEM 0.0024).

Although we obtain a modest (7%) reduction in variance

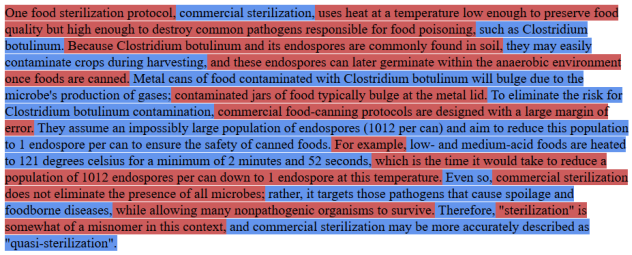

Figure 4: Example of sentence-fragment representation where the alternating colors signify the different fragments.

with the binary fragment-level highlights and the individual-word highlights, it is likely that these seemingly promising results are meaningless because the model degrees of freedom (235 and 2291, respectively) are larger than the number of subjects in our data set (198).

Because our models have sufficient degrees of freedom that they are not well constrained by the data, we turn to simple models that leverage domain knowledge, specifically, our knowledge of which sentence in the text contains the critical information for a given quiz question. We tested the correlation between highlighting a critical sentence and improved performance on the corresponding quiz question. Analyzing the fill-in-the-blank (FIB) and multiple choice (MC) questions separately, a two-tailed matched-sample t-test indicates that MC quiz scores are significantly higher for those who highlighted the critical sentence than for those who did not (0.74 versus 0.63, $t(11) = 4.05$, $p = 0.002$, $d = 0.73$). A marginal effect in the expected direction was also found for FIB by those who highlighted the critical sentence versus those who did not (0.34 versus 0.29, $t(11) = 2.034$, $p = 0.067$, $d = 0.23$).

We then built linear regression models to determine how the conditional analysis of the previous paragraph translates to predictive model performance. Models were built predicting specific quiz question accuracy from the corresponding critical sentence, separately for MC and FIB and for each of the 12 questions. Models were evaluated using 10-fold cross validation. Averaging across the 12 questions, the normalized loss is 0.99 for MC, ranging over questions from 0.94 to 1.02, and the normalized loss is 1.00 for FIB, ranging over questions from 0.95 to 1.03.

Although pre-selecting the critical passage elements does not appear to boost prediction, we are optimistic, based on the conditional probability analysis, that with additional data, our models will begin to reveal dependencies.

In conclusion, none of the linear models are particularly promising. Although two of the models do seem to predict some variance in the test scores—models utilizing the binary fragment-level and the word-level representations—one has to be cautious in reaching a positive conclusion given the large number of models we constructed.

3.1.2 Nonlinear models

We also evaluated nonlinear regression models, specifically neural networks. The neural networks had one or two hid-

den layers with tanh activation functions and an output layer with a single sigmoid unit to represent the normalized test score prediction. The nets were trained by the Adam optimizer to minimize the mean square error between the normalized quiz score and the prediction, with an initial learning rate of 0.001 and batch size equal to 20% of the size of the training set. All weights were initially drawn using Xavier initialization. A validation set was created from 10% of the supplied training data, which was used to stop training after the normalized error on the validation set plateaued after 50 epochs. Model hyperparameters (see table below) were chosen by a grid search in the inner cross validation loop. The regularizers include an L2 weight penalty on the input-to-hidden weights and dropout on the nodes in the hidden layers.

Grid Search	
Hyper Parameter	Values
Dropout rate	0, 0.5
Hidden layer 1 size	5, 10, 15, 20
Hidden layer 2 size	0, 5, 10, 15
L2 regulariz. relative learn rate	0, 0.25, 0.5, 0.75, 1

For each highlight representation (sentence-level, sentence-fragment, individual words), we found the best hyperparameters over the grid search and evaluated the models using 10-fold cross validation. We present the results of each of these networks in Table 3. Unfortunately, none of these models outperformed the baseline.

We hypothesized that there might be information to leverage by predicting performance on individual questions rather than their sum (the total quiz score). We therefore built neural net models with outputs that represent the individual questions, with two output units for each of the 12 questions. The target tuple (0,0) represents neither fill-in-the-blank (FIB) nor multiple-choice (MC) response correct; (0,1) represents FIB incorrect but MC correct; (1,0) represents FIB correct but MC incorrect; and (1,1) represents both FIB and MC correct. The logic of this coding scheme is that the first bit indicates strong knowledge of the answer and the second bit indicates at least weak knowledge.

The training and evaluation process was the same as the neural networks that predict on overall quiz score, with the normalized loss an expectation over the 24 outputs. We evaluated networks for each of the highlight representations, and Table 3 lists the results. Unfortunately, no predictions were better than baseline.

4. DISCUSSION

If you pick up any textbook in a used bookstore, you'll be surprised if it isn't marked up with student annotations and highlights. Students seem compelled to highlight because they believe it supports learning. Our goal was to leverage this compulsion to better understand what students are learning from their textbooks. We hypothesized that a learner's choice of material for highlighting could differentiate among individuals and predict comprehension. We constructed a wide range of models that use the specific pattern of highlights to predict subsequent quiz performance and specific quiz answers, yet we failed to obtain strong support for our hypothesis.

The most generous interpretation of our modeling effort is that when highlights are represented at a fine-level of granularity—sentence fragments or individual words—linear models can predict about 6% of the variability in quiz score. It's difficult to explain why the linear models (Table 2) outperform the nonlinear models with the same input representation, but perhaps we are not successfully controlling for overfitting of the more complex models. The variance in model predictions across cross-validation folds is an indication that the models are perhaps still too flexible and would benefit by stronger regularization.

The present experiment had several sources of uncontrolled variability that, in retrospect, should have been taken into account.

- We neglected to ask participants about their familiarity with biology and we did not exclude participants based on their knowledge. Prior knowledge could be a significant uncontrolled factor. In subsequent experiments, it would be sensible to screen participants based on whether they have had a biology class in the past three years.

- The randomized order of quiz questions influences the interval of time for which knowledge must be retained. For example, if the first quiz question is on the third passage of text, then the lag between reviewing that passage and the quiz question is just a matter of seconds. It would be more sensible to present the quiz questions in order by section and to randomize the order within a passage.

- In the present experiment, participants had little idea of what the quiz would entail until they completed the initial reading and review stages of all three passages. We suspect that participants may highlight in a more informed manner if they can better anticipate what is to come in the experiment. Thus, we might have included in the instructions a sample paragraph and several typical exam questions.

- We encouraged participants to highlight, but we did not ask participants whether they ordinarily highlight text as they read. There seems to be individual differences in the proclivity to highlight, and it would be useful to perform analyses of the highlights for the subpopulations who either do or do not ordinarily highlight.

A natural thought for improving predictive models is to encode information about the content of the text and semantic relationships among the individual sentences and phrases. We argue that such encodings will *not* improve our models for the specific experiment we have performed. If our goal was to devise a general passage-independent representation of text, then incorporating such encodings would be critical, but because we have three specific passages and our highlight representation allows for the reconstruction of which

Table 2: Summary of linear regression results

Input Features	Target Output	Mean Normalized Loss	Standard Error of Mean
Total number of sentence-level highlights	Normalized Quiz Score	1.01	0.0029
Total number of words highlighted	Normalized Quiz Score	1.01	0.0028
Binary sentence-level highlights	Normalized Quiz Score	0.99	0.0028
Graded sentence-level highlights	Normalized Quiz Score	1.03	0.0032
Binary sentence-fragment highlights	Normalized Quiz Score	0.93	0.0024
Graded sentence-fragment highlights	Normalized Quiz Score	1.03	0.0032
Word-level highlights	Normalized Quiz Score	0.93	0.0024
Critical-sentence highlight	Corresponding FIB Question Score	1.00	N/A
Critical-sentence highlight	Corresponding MC Question Score	0.99	N/A

Table 3: Summary of neural network results

Input Features	Target Output	Mean Normalized Loss	Standard Error of Mean
Binary sentence-level highlights	Normalized Quiz Score	1.01	0.0030
Graded sentence-level highlights	Normalized Quiz Score	1.00	0.0022
Binary sentence-fragment highlights	Normalized Quiz Score	0.99	0.0026
Graded sentence-fragment highlights	Normalized Quiz Score	1.20	0.0032
Word-level highlights	Normalized Quiz Score	1.03	0.0021
Binary sentence-level highlights	Individual Question Scores	1.00	0.0049
Graded sentence-level highlights	Individual Question Scores	1.00	0.0052
Binary sentence-fragment highlights	Individual Question Scores	1.00	0.0054
Graded sentence-fragment highlights	Individual Question Scores	1.00	0.0053
Word-level highlights	Individual Question Scores	1.00	0.0050

specific sentences, phrases, or words were highlighted, we argue that this representation is sufficient for prediction. For example, if the participant were to highlight all phrases related to thermal death time, we do not need an explicit representation of this concept because the pattern of sentences highlighted contains this information implicitly.

We have ideas for extending the present work with the hope that highlighting might serve as a valuable data source for inferring student knowledge. We mention several key ideas here.

- We explored a variety of highlighting representations in order to capture critical differences among highlighting patterns. However, we are not convinced that all critical differences are captured. Consider the following sentence from one of the passages in the experiment: *Unlike disinfectants, antiseptics are antimicrobial chemicals safe for use on living skin or tissues.* Highlights of this sentence in our data set include:

  - *antiseptics*
  - *antiseptics are antimicrobial chemicals*
  - *antiseptics are antimicrobial chemicals safe for use on living skin or tissues.*

  All three of these highlights are treated the same by the sentence and fragment representations with the binary encoding, but one might imagine that they provide different windows into the student's intentions. The individual word representation does distinguish these patterns, though at the cost of a much larger input and parameter space. The sentence-level and sentence-fragment graded encodings seem to be a sensible intermediate, but we suspect there are other intermediate encodings that would be fruitful to explore.

- One potentially useful source of information would be the detailed time course of reading, i.e., fixation patterns as a function of time, or at least obtaining information on the rate at which sentences are read and when backtracking occurs. In our current experiment, timing information is recorded only when a sentence is highlighted; these data are too sparse to provide a useful representation that can be compared across individuals.

  In order to record better timing information, we have considered conducting the experiment using a small screen e-reader (or a small window on a computer monitor) which necessitates scrolling from one paragraph to the next.

5. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation award EHR-1631428.

6. REFERENCES

[1] John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest* 14, 1 (2013), 4–58.

[2] Robert L Fowler and Anne S Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology* 59, 3 (1974), 358.

[3] Caitlin Mills, Art Graesser, Evan F Risko, and Sidney K D'Mello. 2017. Cognitive coupling during reading. *Journal of Experimental Psychology: General* 146, 6 (2017), 872.

[4] Sherrie L Nist and Mark C Hogrebe. 1987. The role of underlining and annotating in remembering textual information. *Literacy Research and Instruction* 27, 1 (1987), 12–25.

[5] John P Rickards and Gerald J August. 1975. Generative underlining strategies in prose recall. *Journal of Educational Psychology* 67, 6 (1975), 860.

[6] Sam T Roweis and Lawrence K Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* 290, 5500 (2000), 2323–2326.

[7] Connie Rye, Robert Wise, Vladamir Jurukovski, Jung Desaix, and Yael Avissar. 2016. *Biology.* OpenStax.