Towards Predicting Reading Comprehension From Gaze Behavior

Seoyoung Ahn
Stony Brook University
Stony Brook, New York

Conor Kelton
Stony Brook University
Stony Brook, New York

Aruna Balasubramanian
Stony Brook University
Stony Brook, New York

Gregory Zelinsky
Stony Brook University
Stony Brook, New York

Abstract

As readers of a language, we all agree to move our eyes in roughly the same way. Yet might there be hidden within this self-similar behavior subtle clues as to how a reader is understanding the material being read? Here we attempt to decode a reader's eye movements to predict their level of text comprehension and related states. Eye movements were recorded from 95 people reading 4 published SAT passages, each followed by corresponding SAT questions and self-evaluation questionnaires. A sequence of 21 fixation-location (x,y), fixation-duration, and pupil-size features were extracted from the reading behavior and input to two deep networks (CNN/RNN), which were used to predict the reader's comprehension level and other comprehension-related variables. The best overall comprehension prediction accuracy was 65% (cf. null accuracy = 54%) obtained by CNN. This prediction generalized well to fixations on new passages (64%) from the same readers, but did not generalize to fixations from new readers (41%), implying substantial individual differences in reading behavior. Our work is the first attempt to predict comprehension from fixations using deep networks, where we hope that our large reading dataset and our protocol for evaluation will benefit the development of new methods for predicting reading comprehension by decoding gaze behavior.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Reading, Eye tracking, Machine learning

ACM Reference Format:
Seoyoung Ahn, Conor Kelton, Aruna Balasubramanian, and Gregory Zelinsky. 2020. Towards Predicting Reading Comprehension From Gaze Behavior. In *2019 Symposium on Eye Tracking Research and Applications (ETRA '20), June 2–5, 2020, Stuttgart, DE.* ACM, New York, NY, USA, 5 pages. https://doi.org/XXX

1 Introduction

Eye movements have long been assumed to reflect cognitive processes underlying various viewing tasks, such as reading, scene perception, and visual search [Henderson 2003; Rayner 2009]. Research even exists that attempts to infer a person's cognitive state from their eye-tracking data. While most of these studies have focused on predicting the task (e.g., scene viewing vs scene memorization [Boisvert and Bruce 2016; Henderson et al. 2013]), or inferring some characteristic of an individual (e.g., personality [Al-Samarraie et al. 2017, 2018]), few attempts have been made to predict a person's cognitive state during reading, including their level of text comprehension.

Underwood et al. [Underwood et al. 1990] is one of the first studies predicting reading comprehension from eye movements, and they found that fixation duration was a predictor of comprehension level (high vs low). However, they trained and tested their classifier on the same dataset; generalizability of their method to an unseen dataset is therefore unknown. Other studies [Augereau et al. 2016; Lou et al. 2017] succeeded in decoding reading behavior to predict an individual's literacy skill (80.3% of accuracy), but literary skill here was a score from a separate language test, not measured for the text that was read while the eye movements were recorded. A recent work [Makowski et al. 2018] focused on predicting comprehension based on reading behavior on the text read. This is the same as our goal. Whereas this work failed to make meaningful predictions, it is an open question whether the application of different features and computational methods to this very difficult problem may produce more optimistic results.

In the present study, we explore methods for classifying comprehension related variables for a reader, such as their level of overall text comprehension, individual passage comprehension, perceived difficulty of reading, and whether the reader's first language is English. High-quality fixation data were collected from 95 participants reading four SAT practice passages, which was followed by responses to comprehension questions and self-evaluation questionnaires for each text passage. We compare the classification performance from the two most widely used neural networks, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs), which enable the learning of complex and dynamic features from raw eye-movement data. Details about the methods are described in Section 3.

Using these networks as classifiers, our main goal is to predict a person's comprehension during reading from only their fixation behavior. To the extent that these predictions are successful, this

ETRA '20, June 2–5, 2020, Stuttgart, DE

ACM ISBN XXX…$15.00
https://doi.org/XXX

Table 1: Descriptive statistics for conventional eye-movement measures in comprehension, reading difficulty, and first-language conditions.

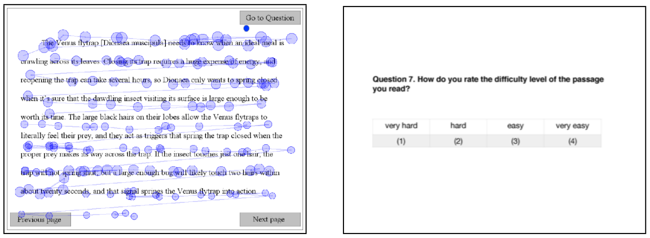
		Overall Comprehension		Passage Comprehension		Reading Difficulty		First Language	
Eyemovement Measures[1]		Low	High	Low	High	Low	High	Non-English	English
Average Fixation Duration (ms)	Mean	214.64	200.91	213.78	204.49	204.72	214.87	219.11	203.31
	SD	29.34	29.75	30.86	29.40	29.22	31.31	32.59	27.94
Average Saccade Amplitude (°)	Mean	4.22	4.68	4.29	4.53	4.50	4.32	4.07	4.60
	SD	0.72	1.06	0.77	1.01	0.95	0.86	0.74	0.95
Average Pupil Size	Mean	1586.19	1592.96	1580.79	1594.94	1597.48	1573.46	1593.61	1587.54
	SD	388.97	499.94	395.38	474.10	420.61	489.23	486.01	425.81
Reading Rate (WPM)	Mean	218.60	245.02	217.62	239.82	234.55	224.36	190.09	249.14
	SD	107.28	78.88	112.87	81.83	103.61	77.71	55.55	103.81

[1] Average fixation duration is the average duration in milliseconds (ms) of all fixations on each passage over participants. Average saccade amplitude is the average amplitude in visual angle (°) of all saccades on each passage over participants. Average pupil size is the average of mean pupil size in arbitrary units (default by Eyelink 1000; the number of thresholded pixels on eye tracking camera) for all fixations made during the reading of each passage over participants. Reading rate is the average WPM (the number of words read per minute) on each passage over participants.

method would eliminate the need for readers to manually answer explicit comprehension or self-evaluation questions, which is a time-consuming, laborious, and potentially obstrusive process. Moreover, accurate prediction would make possible the creation of intelligent systems capable of giving instantaneous feedback to readers about their comprehension or mental state, which could significantly advance the quality of education and training. Finally, our work is of interest to cognitive scientists studying the relationship between reading behavior and comprehension, and specifically the identification of reading patterns leading to a good or poor understanding of a concept.

2 SAT Reading Dataset

Here we introduce one of the largest datasets of gaze fixations during reading. It consists of the eye movement data from 95 undergraduate students reading four SAT passages for comprehension, and their responses on comprehension questions and self-evaluation questionnaires (e.g., subjective difficulty of passage). All reading passages and questions were selected from the SAT practice set from their official website: https://collegereadiness.collegeboard.org/sat/practice. The dataset is publicly available at https://github.com/ahnchive/SB-SAT and we encourage people to download it to develop their own models to predict reading comprehension. A detailed description of the data collection procedure and descriptive eye-movement results is provided below.



The Venus flytrap [Dionaea muscipula] needs to know when an ideal meal is crawling across its leaves. Closing its trap requires a huge expense of energy, and reopening the trap can take several hours, so Dionaea only wants to spring closed when it's sure that the dawdling insect visiting its surface is large enough to be worth its time. The large black hairs on their lobes allow the Venus flytraps to literally feel their prey, and they act as triggers that spring the trap closed when the proper prey makes its way across the trap. If the insect touches just one hair, the trap will not spring shut, but a large enough bug will likely touch two hairs within about twenty seconds, and that signal springs the Venus flytrap into action.

Question 7. How do you rate the difficulty level of the passage you read?

very hard	hard	easy	very easy
(1)	(2)	(3)	(4)

Figure 1: Examples of text and questions used in the reading comprehension experiment. Left: A page of text for reading overlaid with fixations, Right: Self-evaluation question page.

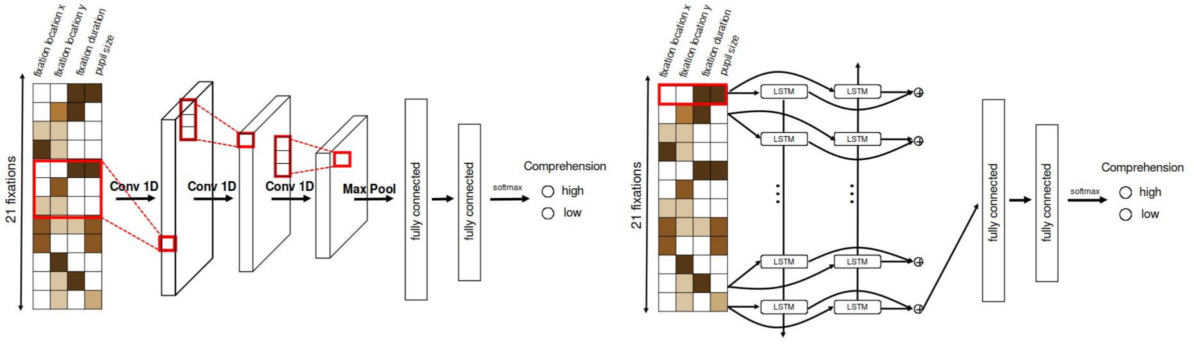
2.1 Procedure and Apparatus

After calibration (average error < .9°, maximum error < 1.5°), participants were instructed to read text pages from a passage, with each page fully displayed one at a time on the computer screen, for the purpose of answering comprehension questions that would follow. Each reading page was presented on a 19-inch flat-screen CRT ViewSonic SVGA monitor (screen resolution of 1024 × 768 pixels, refresh rate of 100 Hz). The screen subtended $30^\circ \times 22^\circ$ of visual angle, making the width of three characters spanning approximately 1°. Participants were allowed up to five minutes to read a passage, but there was no time constraint in answering questions. The order of presentation of four reading passages was counterbalanced by Latin square design. Eye position during reading was recorded using an EyeLink 1000 (SR Research) sampling at 1000 Hz, and gaze coordinates were parsed into fixations using the default Eyelink algorithm having a velocity threshold of $30^\circ/sec$ and an acceleration threshold of $8000^\circ/sec^2$. Calibration drift was checked before reading every page, and recalibration was performed if required.

2.2 Descriptive Results

In Table 1, we report average eye-movement measures for different levels and variables related to comprehension: Overall comprehension, Passage comprehension, Reading difficulty, and First language. Overall comprehension is an individual's comprehension score averaged over all four passages, presumably also reflecting an individual's language proficiency. The level of overall comprehension is defined as "high" if the score is higher than the median (55%), and "low" for the rest. Passage comprehension is an individual's comprehension score for each passage. Similarly, passage comprehension scores higher or equal to the median (60%) are defined as "high" level, while the other scores are grouped as "low" level. For reading difficulty, a high (hard) level was assigned if participants rated the passage as "very hard" or "hard", and a low level was assigned when the passage was rated as "very easy" or "easy". Lastly, we grouped the data based on whether or not the participant reported being a native speaker of English.

Compared to the low-level comprehension group, two-tailed t-tests revealed that people having a high level of overall comprehension were significantly faster and more efficient in their reading



(a) CNN

(b) RNN

Figure 2: General pipeline for network models used in this study. Each layer is followed by Rectified Linear Unit (ReLU) activation, and dropouts are used before each fully-connected layer again excluding the last. Details of each model architecture are described in Section 3.2

behavior; they made shorter duration fixations, t(378) = 4.52, larger saccade amplitudes, t(378) = -5.02, and had a faster reading rate, with all ps < .001). The same trend was observed for passage comprehension in reading rate, t(378) = -2.22, p<0.05, fixation duration, t(378) = 2.95, p<0.001, and saccade amplitude, t(378) = -2.49, p<0.05. We also found that people who rated the passage as difficult tended to read more slowly and carefully, producing fixations having longer durations and saccades having smaller amplitudes, although a significant difference was only observed in fixation duration, t(378) = -3.12, p<0.05. Lastly, people who reported speaking English as their first language generated shorter duration fixations, t(378) = 5.03, larger saccades, t(378) = 4.81, and had a higher reading rate, t(378) = -5.77, compared to people who identified themselves as being a non-native English speaker, with all corresponding ps < .001. Mean pupil size, however, did not differ significantly across levels for the variables considered here. Collectively, these results show that high and low levels of text comprehension were accompanied by different patterns of average eye-movement behavior during reading.

3 Comprehension Prediction

Here we develop a classifier that decodes a reader's eye movements to predict their comprehension-related states. These are: Overall comprehension, Passage comprehension, Reading difficulty and First language, as defined in Sec. 2.2. Models were trained on training datasets; this training ended based on the results from validation datasets; and prediction accuracy is reported based on results from test datasets. Our specific training/validation/test splits of the data are described in Sec. 3.3 in detail.

3.1 Data Slicing and Model Input

To obtain many samples of reading behavior to use for model training, we grouped the eye-movement data into windows of 21 consecutive fixations, where windows were constrained not to overlap with each other. However, in order to be used for model training these windows must be labeled for ground truth comprehension, which would require the annotation of over 10k windows. Previous studies addressed this problem by having people manually label each fixation within a narrow window of time [Biedert et al. 2012; Ishimaru et al. 2017], but the size of our dataset made this impractical. Instead, we had each window inherit the label at the passage level. We simply assigned a ground truth label to each window based on the label of the passage from which the window came, e.g., if a reader has a high comprehension level for a particular passage, all windows of that reader extracted from that particular passage are labeled as high. Note that our method therefore assumes a sort of consistency in scale for an individual reader; that their reading at a local scale (e.g., sentence level) is the same as their reading at a more global scale (passage level). Four oculomotor features from each window were normalized and used for input data to the network: x fixation location, y fixation location, fixation duration, and pupil size, making the final input dimension: 21×4. We realize that other oculomotor features or slicing of the data could have been explored, but this is beyond the scope of our current question.

3.2 Model Architecture

Convolutional Neural Network (CNN). The architecture of the CNN used in this study is shown in Figure 2 (a). The network has three one-dimensional convolutional layers and two fully-connected (FC) layers. For the convolutional layers, 40 filters with size 3 and stride 1 are used without zero-padding. The resulting feature maps are fully connected to 50 units, and these are again fully connected to 20 units. Finally, the last softmax layer has two output units corresponding to the two levels of a predicted variable (e.g., high or low level of comprehension).

Recurrent Neural Network (RNN). The architecture of the RNN used in this study is shown in Figure 2 (b). Our RNN models consist of two Bidirectional LSTM (Long Short Term Memory) layers [Graves and Schmidhuber 2005], followed by two FC layers. For each LSTM cell, 25 hidden units are used. The same design is applied for the rest of the network (e.g., FC layers and output layer) as described above for the CNN.

3.3 Evaluation

Three types of prediction were performed in this study: prediction of new reading fixation windows (samples), prediction of fixation windows on new passages, and prediction of fixation windows on new subjects. Each method required a different split of the training/validation/test data, and all evaluations were done in terms of prediction accuracy on our test dataset. We report the results of our CNN and RNN models, as well as a shallow-learning regression model and Null model baselines. The Null model is our lower bound, and simply outputs labels as the most frequent class in the dataset. Refer to the Sec. 2.2 for detailed descriptions of the predicted variables.

Prediction on New Reading Windows. New reading windows are those that were not seen by the model during training, and here we report predicted comprehension levels for these windows. A total of 11,548 samples were randomly split into the proportions of 60%/20%/20% for training/validation/test datasets, respectively. As can be seen in Table 2 (a), Overall comprehension was best predicted at 65% by our CNN model, which is 11% higher than null accuracy. Passage comprehension was predicted equally well by both the CNN and RNN models (62% accuracy), but this was marginally above Null model accuracy (61%) and therefore not meaningful. As for predicting Reading difficulty, all models were at or near Null accuracy (67%). First language was best predicted by the CNN, which attained 71% accuracy (5% > Null accuracy).

Prediction on New Passages. Next, we tested whether model predictions could generalize to fixation windows from new passages (e.g., predicting a reader's comprehension for one unseen passage after training with that person's reading behavior from the two other passages).We performed a 2/1/1 passage-level split the data, meaning that 2 passages were used for training, 1 for validation, and 1 unseen passage for testing. Different random splits of the passages were used for different readers. As shown in Table 2 (b), Overall comprehension was predicted at 64% accuracy by both the CNN and RNN models, surpassing Null model accuracy by 11%. As for Passage comprehension and Reading difficulty prediction, all models performed poorly, below null accuracy. First language was best predicted at 69% by the CNN (3% > Null accuracy).

Prediction on New Readers. Finally, we tested whether model predictions would generalize to fixation windows from new readers. Data samples were split to reflect 57/19/19 readers used for training, validation, and testing, respectively, meaning that comprehension was predicted for 19 unseen readers after training with the data from 57 different readers. However, as can be seen in Table 2 (c), all models performed poorly, below or near Null accuracy.

4 Conclusions and Future Directions

In this study we explored methods for classifying a reader's level of comprehension, specifically Overall text comprehension and Passage-level comprehension, as well as the comprehension-related variables of perceived Reading difficulty and whether the reader's First language was English. Our deep network models (CNN/RNN), given windows consisting of 21 consecutive fixations made during the reading of SAT passages, predicted overall reading comprehension at a 65% level of accuracy. This level is not astounding, but given a null accuracy of 54%, an 11% increase in comprehension prediction may be important to the Education, Learning, and Training research communities. Moreover, the prediction success of these models generalized from old to new passages (best: 64%), but did not generalize well from old to new readers (best: 41%). This pattern argues for the targeting of reading-based methods for identifying poor comprehension to individual readers. Models predicting the other variables that we considered performed no better than null accuracy, except for the CNN being slightly better than the Null model in sample and passage-wise First language prediction.

Table 2: A comparison of model accuracy (Sec. 3.2) on the predicted variables (Sec. 2.2). Best predictions are in bold text.

Predicted Variable	Null Acc.	Regression	CNN	RNN
Overall Comprehension	0.54	0.57	**0.65**	0.64
Passage Comprehension	0.61	0.6	**0.62**	**0.62**
Reading Difficulty	**0.67**	0.64	0.6	**0.67**
First Language	0.66	0.63	**0.71**	0.69

(a) Prediction on New Reading Windows

Predicted Variable	Null Acc.	Regression	CNN	RNN
Overall Comprehension	0.53	0.57	**0.64**	**0.64**
Passage Comprehension	**0.61**	0.49	0.56	0.53
Reading Difficulty	**0.64**	0.6	0.61	0.62
First Language	0.66	0.64	**0.69**	0.68

(b) Prediction on New Passages

Predicted Variable	Null Acc.	Regression	CNN	RNN
Overall Comprehension	**0.5**	0.37	0.4	0.41
Passage Comprehension	0.59	0.53	0.56	**0.6**
Reading Difficulty	**0.77**	0.55	0.65	0.67
First Language	**0.77**	0.59	**0.77**	0.75

(c) Prediction on New Readers

Our first attempt to predict reading comprehension level using deep networks left much room for improvement. One major limitation of our approach is that we assigned ground truth labels to local fixation windows based on the passage-level labels. We did this for practical reasons, but addressing this problem is a direction for future work. One promising solution from computer vision is to learn a weighting of local features that optimizes global prediction [Wei et al. 2016; Wei and Hoai 2016], and, indeed, a very recent study [Kelton et al. 2019] used this method to build a reading vs skimming detector that did not require explicit ground truth labels for local fixation windows. In future work we also plan to explore more abstract feature inputs, such as saccade amplitude, velocity, angle, and regression. We will also explore the inclusion of lexical features (e.g., lexical frequency or word length), which we expect will improve the performance of our predictive models given the demonstrated importance of lexical factors in the control of reading behavior [Rayner 1998].

We conclude by agreeing with Makowski et al. [Makowski et al. 2018], predicting comprehension from reading fixations is a difficult problem. Our work also highlights the distinction between statistical testing and prediction; prediction was often impossible despite significant differences in the average eye-movement measures between comprehension levels. Nevertheless, we showed that limited success is achievable, and we hope that our methods and large reading-fixation dataset will fuel additional research into comprehension prediction. Given the accelerating pace that eye-tracking paradigms are reaching into diverse contexts, even modest increases in the ability to predict comprehension from reading behavior might have widespread impact.

Acknowledgments

This work was supported by the United States National Science Foundation under grant CNS-1718014.

References

Hosam Al-Samarraie, Atef Eldenfria, and Husameddin Dawoud. 2017. The impact of personality traits on users' information-seeking behavior. *Information Processing & Management* 53, 1 (2017), 237–247.

Hosam Al-Samarraie, Samer Muthana Sarsam, Ahmed Ibrahim Alzahrani, and Nasser Alalwan. 2018. Personality and individual differences: the potential of using preferences for visual stimuli to predict the Big Five traits. *Cognition, Technology & Work* (2018), 1–13.

Olivier Augereau, Hiroki Fujiyoshi, and Koichi Kise. 2016. Towards an automated estimation of English skill via TOEIC score based on reading analysis. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 1285–1290.

Ralf Biedert, Jörn Hees, Andreas Dengel, and Georg Buscher. 2012. A robust realtime reading-skimming classifier. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 123–130.

Jonathan FG Boisvert and Neil DB Bruce. 2016. Predicting task from eye movements: On the importance of spatial distribution, dynamics, and image features. *Neurocomputing* 207 (2016), 653–668.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5-6 (2005), 602–610.

John M Henderson. 2003. Human gaze control during real-world scene perception. *Trends in cognitive sciences* 7, 11 (2003), 498–504.

John M Henderson, Svetlana V Shinkareva, Jing Wang, Steven G Luke, and Jenn Olejarczyk. 2013. Predicting cognitive state from eye movements. *PloS one* 8, 5 (2013), e64937.

Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise, and Andreas Dengel. 2017. Towards reading trackers in the wild: detecting reading activities by EOG glasses and deep neural networks. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 704–711.

Conor Kelton, Zijun Wei, Seoyoung Ahn, Aruna Balasubramanian, Samir Das, Dimitris Samaras, and Gregory Zelinsky. 2019. Reading Detection in Real-time. In *In 2019 Symposium on Eye Tracking Research and Applications (ETRA '19)*.

Ya Lou, Yanping Liu, Johanna K Kaakinen, and Xingshan Li. 2017. Using support vector machines to identify literacy skills: Evidence from eye movements. *Behavior research methods* 49, 3 (2017), 887–895.

Silvia Makowski, Lena A Jäger, Ahmed Abdelwahab, Niels Landwehr, and Tobias Scheffer. 2018. A Discriminative Model for Identifying Readers and Assessing Text Comprehension from Eye Movements. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 209–225.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology* 62, 8 (2009), 1457–1506.

Geoffrey Underwood, Alison Hubbard, and Howard Wilkinson. 1990. Eye fixations predict reading comprehension: The relationships between reading skill, reading speed, and visual inspection. *Language and speech* 33, 1 (1990), 69–81.

Zijun Wei, Hossein Adeli, Minh Hoai Nguyen, Greg Zelinsky, and Dimitris Samaras. 2016. Learned Region Sparsity and Diversity Also Predicts Visual Attention. In *Advances in Neural Information Processing Systems*. 1894–1902.

Zijun Wei and Minh Hoai. 2016. Region ranking SVM for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2987–2996.