Towards Robust Gaze-Based Objective Quality Measures for Text

Ralf Biedert*
Andreas Dengel†
German Research Center for Artificial Intelligence

Mostafa Elshamy‡
German University in Cairo

Georg Buscher§
Microsoft Bing

Abstract

An increasing amount of text is being read digitally. In this paper we explore how eye tracking devices can be used to aggregate reading data of many readers in order to provide authors and editors with objective and implicitly gathered quality feedback. We present a robust way to jointly evaluate the gaze data of multiple readers, with respect to various reading-related features. We conducted an experiment in which a group of high school students composed essays subsequently read and rated by a group of seven other students. Analyzing the recorded data, we find that the amount of regression targets, the reading-to-skimming ratio, reading speed and reading count are the most discriminative features to distinguish very comprehensible from barely comprehensible text passages. By employing machine learning techniques, we are able to classify the comprehensibility of text automatically with an overall accuracy of 62%.

CR Categories: I.7.m [Computing Methodologies]: Document and Text Processing—Miscellaneous; J.5 [Computer Applications]: Arts and Humanities—Literature

Keywords: objective text measures, eye tracking, machine learning

1 Introduction

Today digital sales of books already surpass the numbers for printed editions [Miller and Bosman 2011] and this trend is probably going to continue in the future. At the same time we see eye tracking devices become miniaturized and integrated into tablets and notebooks and there is a chance that these devices may become mainstream as web cams are already integrated in screens and smartphones.

And while in many parts of the digital media world the possibilities to collect and generate end-user feedback have already been embraced, for example in advertisement [Duchowski 2002] or computer games where the player's level progress is being measured to identify difficult parts [Kennerly 2003], [Hunicke 2004], in this paper we investigate how similar means could be developed on text.

Our vision is that in the future a subsequent edition of a book could be optimized by the objective and implicitly gathered reading behavior of the readers of the previous edition, hence assisting authors and editors by providing them new measures and insights into their reader's behavior and progress in the text.

*ralf.biedert@dfki.de
†andreas.dengel@dfki.de
‡mostafa.el-hosseiny@student.guc.edu.eg
§georgbu@microsoft.com

ETRA 2012, Santa Barbara, CA, March 28 – 30, 2012.
© 2012 ACM 978-1-4503-1225-7/12/0003 $10.00

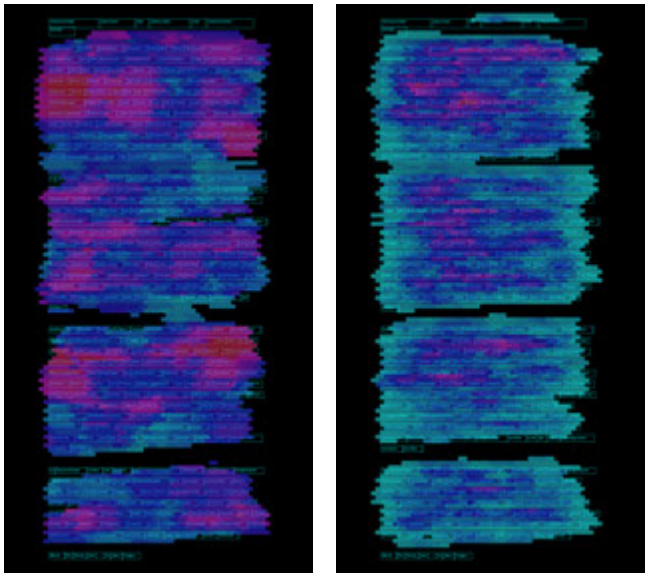

Figure 1: *Sample output of two generated layers for the same document read by seven users. The left image reflects the average relative reading speed at the given passage (red regions have been read very slowly by most readers), the right image depicts the relative reading count (red regions have been read many times by most readers). The upper left hot spot in the left image for example nicely matches with an erroneous sentence one of our authors produced.*

While we are aware that a *straightened* or coherent text does not necessarily improve its usefulness for every reader (compare for example [Kamalski et al. 2008]) and that there is already plenty of research focusing on the inherent textual readability [DuBay 2004], we nonetheless believe that the aggregated knowledge about how it was *actually* interacted with will yield many highly interesting insights. Yet, such an aggregation of reading data is even more intriguing, as the improvements that might be gained through it would be for the good of all readers, not merely the ones wealthy or willing enough to use a gaze-aware reading device.

The remainder of the paper is organized as follows: In Section 2 we will provide details of our system to aggregate and analyze the reading data of multiple users, how we deal with inaccuracies of gaze data, and which features we consider. Section 3 outlines an experiment we conducted, in which six high school students were asked to compose texts, which were subsequently read and rated by an independent group of seven students. In Section 4 we present our findings and then provide an outlook onto future work in Section 5.

2 Approach

Given a text and eye tracking data of many readers we have two principal goals. The first is to deal with inherent eye tracking error and its impact on aggregation. The second is to investigate the prediction power of various individual aggregation and evaluation algorithms with respect to their potential to analyze and predict certain *issue areas* or hot spots within the text, and to provide general

quality estimates about the text in general. We deliberately defined the term *issue* or aspect so far only weakly, since it depends somewhat on the specific circumstances. A very simple aspect might be the relative amount of users that actually read—or ignored—a certain passage. Closely associated might be the question whether a passage was relevant in the context of a document or not [Buscher 2010]. A considerably more complex measure would be the percentage of users for which eye tracking behavior correlating with *comprehension problems* has been observed—a matter we eventually try to address in this study nonetheless.

While there is plenty of research on how texts are being read [Rayner 1998], our approach differs from the traditional psychological works of computing individual gaze measures on word and character level on the topic in several ways. First, we are interested in a fully automated analysis of recorded gaze data. Second, any algorithm should inherently be able to deal with inaccuracies to which gaze data is subject *in the wild*. Third, while our approach often cannot use existing psychological findings on the matter on an individual level straight away (due to the aforementioned inaccuracies), we nonetheless strive to find measures that are in line with established research on the matter of eye movements and comprehension [Lenzner et al. 2011]. With these issues in mind we now present our method to address them.

2.1 Preprocessing

In principal we consider HTML documents. When a user reads in a browser the raw gaze data along with the page's geometry, i.e., word and image bounding boxes and content, is being recorded using the Text 2.0 framework [Biedert et al. 2010]. For each of the individual users the fixations in screen coordinates are eventually converted to fixations in document coordinates and subsequently used for the computation of the following measures. The fixations are detected by filtering the raw data using a virtual median filter and applying a $25px$, $100ms$ dispersion window, see [Biedert et al. 2012] for details.

In order to deal with inaccuracies of eye tracking measurements we introduce a two-stage grid as the principal unit of data processing, compare Figure 2. The base layer of the grid G consists of tiles $t \in G$ with sizes approximately equivalent to the underlying text. Each of these base tiles in turn has a consideration area $a(t)$, which extends around its sides by the expected amount of eye tracking error, ensuring that all base tiles take into account all gaze data that might have contributed to the text covered by the tile[1]; and eventually all tiles overlapping an individual word are associated to it.

2.2 Measures and Algorithms

For each of the tiles t we now compute various reading-related features which will later on serve as the input for an automated feature selector. The selector then picks their most predictive combination, in our case measures about comprehensibility. However, since many measures are highly individual we need a *fair* way to aggregate these for several users. Therefore, for all of the presented features, we do not consider the values themselves, but rather their percentile per a single user and document.

Average Fixation Duration Given a tile t and the set of a user's document fixations F we consider all fixations $f \in F$ which occurred in $a(t)$. We compute their average time and each tile is assigned the relative ranking $adur(t)$, ranging from 0 for the tiles with the least average time observed to 1 for the tiles with the highest average, with respect to this document and this user, compare [Rayner 1998] via [Hyrskykari 2006], esp. $t(w_x)$ therein.

[1] Since in the scope of this experiment all documents are rendered the same way we decided against putting the consideration areas around the words themselves (then somewhat similar to magnetic [Hyrskykari 2006] boxes or lines) in order to improve the visualization, however, in the long run considering words as the fundamental grid elements G will ease their aggregation when different rendering positions are an issue.

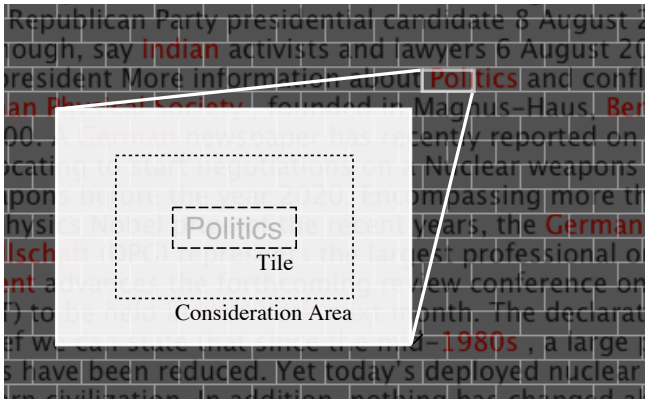

Figure 2: *Sample grid and consideration area. The grid covers the page, including images and other elements, with a tile size approximately similar to the word size, while the consideration area extends to each side by the estimated tracking error.*

First Fixation Duration Similar to the average fixation duration $adur$ we merely consider the duration of the first fixation $fdur(t)$ within the consideration area of a single tile, compare $\mathit{ff}(w_x)$ in [Hyrskykari 2006].

Reading / Skimming Classification Based on a reading-skimming classifier [Biedert et al. 2012] $cls(t)$, we compute for each tile to what extent the saccades intersecting it suggest reading or skimming behavior, ranging from 0 to 1.

Reading Count The number of passes $cnt(t)$ classified as reading intersecting the consideration area of the tile. The idea is that difficult passages are likely to be read again, compare for example [Lenzner et al. 2011].

Reading Saccade Length The minimum length saccade $len(t)$ of all saccades classified as *reading* intersecting the consideration area of the tile, motivated by the idea that problematic passages should be read more *slowly*.

Regression Saccade Sources / Targets The number of times a regression saccade started $src(t)$ in the consideration area of the tile and the number of times a regression saccade ended $tgt(t)$ in the consideration area of the tile.

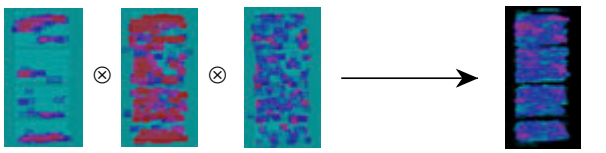

Figure 3: *For each of the previously described features we generate an individual tile layer that might contribute to an aspect like comprehensibility. Using a machine learning algorithm and the recorded ground truth we then try to learn the best possible feature combination to reflect the aspect.*

2.3 Machine Learning

In this paper, we focus on common comprehension difficulties as the issue or aspect to identify, and therefore, we need to find a mechanism that can be trained to do so. Hence, we consider the described features as our main input and employ a machine learning algorithm to find their most predictive combination with respect to a given aspect as measured by human judges, compare Figure 3. For the purpose of this study we perform an SVM classification with parameters C and γ optimized in grid search, using the attributes described in the previous section for each tile. The actual ground truth (i.e., data labeled by human readers according to their comprehensibility) is generated by the following experiment.

3 Experiment

The number of reasons why a text is being read is quite large and in turn it also affects *how* it is read. Thus we have to focus on a specific setting in this work. Since we believe that in the long term such a technology would be interesting especially in the education domain we decided to perform an experiment with high school students in a reading comprehension setting. However, as the deliberate manipulation of texts can yield some (admittedly interesting) problems we decided against manipulating texts on our own, but instead asked a peer group of students to compose novel texts for us. Hence, five senior German high school students were asked to each write texts about a topic of their choice. The only limitations were that the documents should have a sufficient length of about two screen pages, and that they should write the texts on their own and submit their first draft without having proofread them. By using this method we could collect five documents on various topics (e.g., a work diary, a comparison of Italian and German schools, do-it-yourself-car repairs) which fortunately also contained a number of unstaged problems and we found various spelling errors in addition to grammatical problems, ambiguities and obscurities.

We then invited seven other participants, also senior German high school students in the age of 16 to 19 years, two of them female and two of them wearing vision aids. They were told to read the texts (on a Tobii 1750 eye tracker) as part of an exam preparation and they were informed that they would be surveyed about the presented topics afterwards. Each text was immediately followed by a comprehension test. In the end, we presented them another survey in which they were asked to rate selected fragments (sentences or small paragraphs) of the preparation texts on a school grade from 1 (best possible rating) to 6 (worst possible rating), according to how comprehensible these texts were. Each student rated the same six questions per document. In general we tried to balance three of the most comprehensible items (i.e., straightforward and error-free passages) of each document with three of the most problematic items (e.g., grammatical mismatches, ambiguous pronouns or spelling mistakes) for the rating. Also, the order in which the actual documents were read and answered and the order in which the sections were rated was randomized.

Overall, each pass lasted for about 45 minutes including calibration and in the end we obtained 30 ratings for six of our seven users each (180 ratings in total; one user aborted the experiment due to time constraints). These ratings then served as the basis for the training and test sets used in the subsequent evaluation.

4 Evaluation

Analyzing our participants' ratings we find that many of the ratings are unexpectedly high compared to our (the authors) own judgment[2]. There are 91 ratings with a 1 (very comprehensible) label,

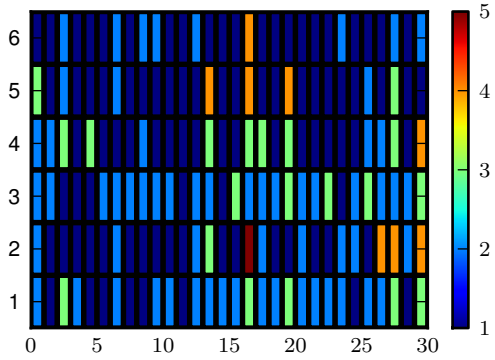

Figure 4: *Six of our seven users rated 30 select fragments for comprehensibility. The aggregated values for each fragment served as the ground truth for the subsequent machine learning step.*

60 ratings with a 2 (comprehensible), 20 ratings with a 3 (mostly comprehensible, took some time), 8 ratings labeled as 4 (probably unclear), 1 rating with a 5 (some parts definitely unclear) and no 6 (unintelligible) label. After aggregating the ratings by computing the average for each fragment in the ground truth and rounding the result we ignored passages with an aggregated rating of 2, merged passaged aggregated to 3 and 4 and decided to consider only these two extreme rating labels (from now on *good* and *bad*) for the classification training. This leaves 17 rated fragments out of the original 30 items.

Each rated item i in the ground truth is mapped to a number of tiles $R_i = \{t_{i1} \dots t_{in}\} \subset T$ where each tile includes values for each reading measure and has a rating equivalent to that of the item it belongs to, generating a total of $\sum R_i = 3578$ tiles. The exact number of elements in each R_i varies according to the geometry of t relative to the text and its intersection with the grid G. These tiles, along with their respective ratings, constitute the training set used to train the classifier. An important factor that affects the actual values included in each tile t is the size of the consideration area $a(t)$. Since the size should correspond to the expected amount of eye tracking error, and since we found the inaccuracy to be rather low in our recording, the area was set to 20px.

Using ground truth with extreme ratings of good and bad, we analyzed the relation between each reading measure for all 3578 (2427 rated as good, 1151 rated as bad) tiles and the rating value, see Figure 5. Calculating the difference between the average values per feature for each rating class gives an overview about which features are estimated to have the most predictive power, and how these features behave in general regarding both classes. See Table 1 for a ranking of features according to this metric.

Next we train the classifier. We balance both classes to a 50:50 distribution by removing 1276 randomly selected *good* tiles and perform a grid search with the RBF-SVM over the parameters C and γ. This was performed once without automated attribute selection and once with automated attribute selection.Training the classifier without automated attribute selection correctly classified 62.12% (1430 out of 2302) of the instances. Performing automated attribute selection before training the classifier predicted reading saccade length, reading / skimming classification, read count and regression saccade targets as the best attribute subset, which is in line with our findings on their statistical significance. The resulting classifier ($C = 3334$ and $\gamma = 0.001$) correctly classified 62.25% ($\kappa = 0.24$) of the instances on 10-fold cross validation.

[2]For example, the (literally translated) sentence "Afterward [sic] they visit the scuola media, which is being attended for three years [missing period] Before they can switch schools [...]" was mostly considered being comprehensible. When asking the students why they had given favorable ratings they stated that the content was comprehensible anyway and they could just read over it.

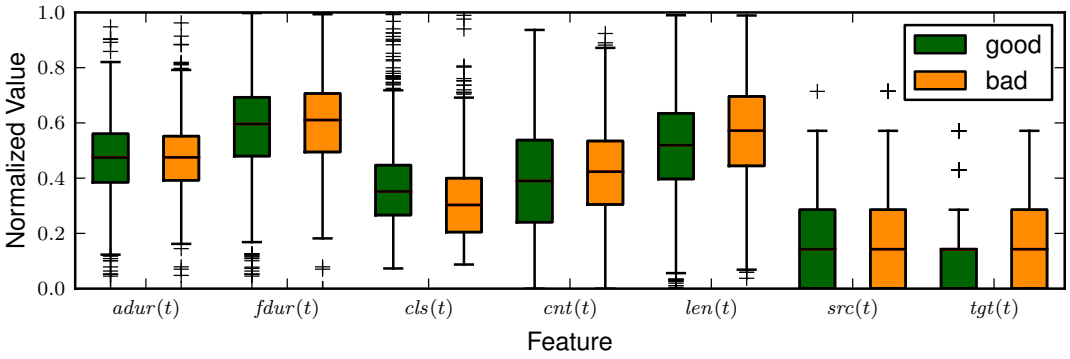

Figure 5: *Box plot for relative aggregated feature values for all our seven considered features and the classes* good *and* bad*. While some features like the saccade target count yield high values others such as the average fixation duration showed no significant difference.*

	Avg. *bad*	Avg. *good*	Δ%	p-KW
Regression Saccade Targets	.168	.126	28.6	.000
Reading-Skimming Class.	.318	.367	14.3	.000
Reading Saccade Length	.567	.509	1.7	.000
Read Count	.416	.385	7.7	.001
Regression Saccade Sources	.145	.138	4.9	.624
First Fixation Duration	.6	.588	2.0	.016
Avg. Fixation Duration	.476	.477	0.2	.754

Table 1: *Average normalized value per feature for all ground truth for the classes* good *and* bad*, the absolute difference in percentages of the maximal possible range, ranked in descending order, and p-values on a Kruskal-Wallis-test.*

5 Conclusion & Outlook

We present a system that is capable of robustly combining gaze data of various users on documents. While it implements a number of features known to correlate with comprehension issues, the employed mechanism can be trained to any sort of human-rated ground truth. In order to assess its predictive power we invited a number of high school students to compose texts which have subsequently been read and rated by a second group of students. Training on the best and worst rated tiles we were able to achieve a 62% classification accuracy in cross validation. Apparently the features reading speed, read-skim classification, regression targets and reading count proved to be the most effective features, while the average fixation duration did not contribute any significant prediction power.

The overall accuracy differs significantly from a guessing level. However, it has not reached a level yet that would make its predictions reliable enough to objectively address text problems based on its output. One issue we encountered in this respect was the - in our opinion - often counterintuitive way the users rated some of the text passages. We assume that an improvement in the specific nature of the questions surveyed will also yield to clearer classes. There are however natural limits to the amount of information that can be asked, and also the timing when this information is asked.

Closely related are two questions regarding the aggregation procedure and its evaluation. For the evaluation we picked a number of text fragments that were commonly rated as good or bad and considered the users' gaze data on them. While for an exploration study like ours this can yield interesting results, the question how these features behave *individually*[3] with respect to comprehensibility (or other aspects) remains highly interesting. Lastly the question what might be the most effective level of granularity (words, lines, paragraphs) for the classification of text *fragments* could also be addressed.

[3]For example, for one user regressions might be a better indicator than for others.

References

BIEDERT, R., BUSCHER, G., LOTTERMANN, T., SCHWARZ, S., MÖLLER, M., AND DENGEL, A. 2010. The Text 2.0 Framework. *Workshop on Eye Gaze in Intelligent Human Machine Interaction*.

BIEDERT, R., BUSCHER, G., HEES, J., AND DENGEL, A. 2012. A Robust Realtime Reading-Skimming Classifier. In *Proceedings of the 2012 Symposium on Eye-Tracking Research & Applications*.

BUSCHER, G. 2010. *Attention-Based Information Retrieval*. PhD thesis, University Kaiserslautern, Kaiserslautern.

DUBAY, W. H. 2004. The Principles of Readability. *Impact Information* (Aug.).

DUCHOWSKI, A. T. 2002. A Breadth-First Survey of Eye Tracking Applications. *Behavior Research Methods, Instruments, and Computers 34*, 4, 455–470.

HUNICKE, R. 2004. AI for dynamic difficulty adjustment in games. *Proceedings of the Challenges in Game AI Workshop, Nineteenth National Conference on Artificial Intelligence.*.

HYRSKYKARI, A. 2006. Eyes in Attentive Interfaces: Experiences from Creating iDict, a Gaze-Aware Reading Aid. *acta.uta.fi*.

KAMALSKI, J., SANDERS, T., AND LENTZ, L. 2008. Coherence Marking, Prior Knowledge, and Comprehension of Informative and Persuasive Texts: Sorting Things Out. *Discourse Processes*, 45, 323–345.

KENNERLY, D., 2003. Better Game Design Through Data Mining, Aug.

LENZNER, T., KACZMIREK, L., AND GALESIC, M. 2011. Seeing Through the Eyes of the Respondent: An Eye-tracking Study on Survey Question Comprehension. *International Journal of Public Opinion Research 23*, 3 (Aug.), 361–373.

MILLER, C. C., AND BOSMAN, J. 2011. E-Books Outsell Print Books at Amazon. *New York Times* (May), B2.

RAYNER, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*.