Predicting Text Readability from Scrolling Interactions

Sian Gooding
Dept of Computer Science and Technology
University of Cambridge
shg36@cam.ac.uk

Yevgeni Berzak
MIT BCS
berzak@mit.edu

Tony Mak
Google
tonymak@google.com

Matt Sharifi
Google
mns@google.com

Abstract

Judging the readability of text has many important applications, for instance when performing text simplification or when sourcing reading material for language learners. In this paper, we present a 518 participant study which investigates how scrolling behaviour relates to the readability of English texts. We make our dataset publicly available and show that (1) there are statistically significant differences in the way readers interact with text depending on the text level, (2) such measures can be used to predict the readability of text, and (3) the background of a reader impacts their reading interactions and the factors contributing to text difficulty.[1]

1 Introduction

There are multiple attributes of a written text that impact how difficult it is to read. This concept is formally known as the *readability* of text, where readability is defined as the sum total of elements within a text that impact a reader's understanding, reading speed and level of interest (Dale and Chall, 1949). Many factors can influence the readability of text, such as the lexical and syntactic complexity, level of conceptual difficulty and style of writing (Xia et al., 2016). For instance, Figure 1 presents an example sentence (b) that has been rewritten to a more readable format shown in (a).[2]

(a) A new scientific study says that global warming might make temperatures rise more than people think.

(b) Temperature rises resulting from unchecked climate change will be at the severe end of those projected, according to a new scientific study.

Figure 1: Example of two sentences (a) and (b) with differing levels of readability. Sentence (b) has a more sophisticated syntactic structure and advanced vocabulary than (a).

Automatically measuring the readability of text has many useful applications. For example, it is used in text simplification (Aluisio et al., 2010), when sourcing reading material for language learners (Xia et al., 2016), in ranking search engine content for dyslexic users (Morris et al., 2018) and to ensure critical consumer information is delivered at an appropriate level (Zou et al., 2019). However, current approaches for measuring readability rely exclusively on linguistic features which do not account for the subjective needs of readers. As a result, traditional readability formulas perform poorly in modeling adult judgements of textual complexity (Crossley et al., 2017).

Furthermore, traditional readability approaches do not work well for online content (Collins-Thompson, 2013). This is due to systems being highly sensitive to noise; requiring well formed sentences and performing poorly on short passages (Collins-Thompson, 2014). Linguistically driven techniques are also language specific and require sophisticated models to extract features. For low-resource languages, such tools and models may be unavailable (Agić et al., 2016).

Achieving a more inclusive assessment of text readability requires collecting subjective measures on multiple levels (e.g., reading interactions and comprehension-based questions) across different genres and populations. Therefore, alternative approaches must be examined to further the understanding of text difficulty and the multiple factors that influence it. In this paper, we introduce such an alternative approach, by using crowdsourcing to collect scroll-based interactions whilst participants read texts at differing levels. We present a 518 participant study, recording the reading interactions

[1] Our dataset is available at https://github.com/siangooding/readability_scroll

[2] Example from the OneStopEnglish dataset by Vajjala and Lučić (2018).

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 380–390
November 10–11, 2021. ©2021 Association for Computational Linguistics

of individuals and obtain statistically significant results showing that reading interactions differ dependent on the textual complexity.

In the following sections, we discuss relevant previous work, outline our data collection and report results from our study. As a preliminary evaluation, we report the statistical significance of reading interactions based on text difficulty, and integrate scroll features into a readability classifier. Finally, we investigate group differences in readability based on the first language of participants. To conclude, in this paper we make the following contributions:

- We release a dataset containing the reading interactions and comprehension scores of 518 participants based on the OneStopEnglishQA dataset (Berzak et al., 2020).
- We show that scroll-based reading interactions can be used to predict the readability of text as well as improve current state-of-the-art approaches.
- We investigate how first language impacts reading interactions and emphasise that the factors contributing to text difficulty vary depending on the target audience.

2 Related Work

Text Readability

Research on assessing the readability of English texts has spanned several decades. The earliest works focused on the construction of readability formulae and metrics (Chall, 1958; Klare et al., 1963; Zakaluk and Samuels, 1988). These measures rely on shallow textual characteristics such as the number of sentences, average length of sentences and average word length. For example, one of the most well known readability scores is the Flesch-Kincaid score (Kincaid et al., 1975). This score takes the average number of words per sentence as well as the average number of syllables as a proxy for syntactic and semantic difficulty. However, there are multiple limitations to readability formulae (Collins-Thompson, 2014). Firstly, these formulae are based on surface characteristics of text, and ignore deeper levels of text processing known to be important factors of readability. Furthermore, these metrics typically assume the text will contain no noise. As a result of this, a number of studies have demonstrated that these metrics are unreliable for web-based content (Si and Callan, 2001; Collins-Thompson and Callan, 2004; Feng et al., 2009).

Due to the limitations of traditional formulae, research in readability assessment subsequently shifted towards machine learning (ML) based techniques. These approaches combine a far richer variety of linguistic features using ML classification algorithms and result in far better performance (François and Miltsakaki, 2012). Early work on statistical readability assessment demonstrated the improvement of these approaches for readability prediction (Si and Callan, 2001; Collins-Thompson and Callan, 2004). Subsequent work focused on the addition of appropriate features, for instance, lexical and grammatical (Heilman et al., 2007, 2008) as well as discourse-based (Pitler and Nenkova, 2008; Feng et al., 2010; Graesser et al., 2011). Systems specifically designed for non-native audiences have also been developed (Feng et al., 2010; Vajjala and Meurers, 2012; Xia, 2019). Additionally, the use of eye-tracking data from both language learners and native speakers has been shown to improve readability assessment models (González-Garduno and Søgaard, 2018). However, many of the aforementioned models require extensive feature engineering, which not only rely on well-formed content, but also depend on multiple resources such as parsers and word-lists.

Implicit Feedback Techniques

Reading on a device, such as a tablet or phone, has predominantly taken the place of traditional formats (Li et al., 2019). Such devices allow access to implicit user feedback by measuring how a user interacts with the text they read. A key advantage of implicit feedback techniques is that they can unobtrusively obtain information by measuring user interactions with a system (Kelly and Teevan, 2003). For instance, Claypool et al. (2001) designed a study to capture mouse and keyboard interactions as implicit measures of interest. They measure the correlation of these implicit features with user ratings and find the time spent and amount of scrolling had a positive correlation with interest. Chen et al. (2021) analyse the factors predictive of English language typing times to investigate effects of linguistic context on language production. Implicit feedback techniques have additionally been shown as useful in information retrieval (Golovchinsky et al., 1999), ranking summaries

(White et al., 2002) and identifying user preferences (Kelly and Teevan, 2003).

In our study, we measure implicit feedback from participant interactions whilst reading. The goal is to produce a set of readability features, using aggregate scroll interactions, that are robust to noise and do not require extensive feature engineering or external resources.

3 Data Collection

We design a controlled reading task and record user reading interactions. We collect responses from participants in the United States and India. This decision was made based on our pilot studies which showed participants from India were much more likely to know English as a second language. The experiment was designed in line with common practises in scroll interaction research (Hinckley et al., 2002). This study represents best efforts in the collection of a novel dataset of reading interactions for text at different levels of readability.

Sampling and Participants

Participants in the study were recruited via a crowdsourcing platform and were based in either India or the US. A range of background information pertinent to language proficiency was collected using a demographic questionnaire which is available on our dataset repository. This included the self-reported English proficiency, native language, hours spent reading per week and highest level of formal education. The design of the questionnaire was informed following guidelines on judging reading ability (Acheson et al., 2008). English was the first language of 66% of participants and a total of 46 distinct first languages were recorded. The second most frequent first language was Tamil, accounting for 19% of participants.

Materials

Investigating how readability impacts user interactions required texts at different levels of difficulty. For this, we use the OneStopEnglishQA dataset (Berzak et al., 2020) which has been designed for the evaluation and analysis of reading comprehension in English. The dataset contains 30 articles from a prior dataset collected for readability assessment and automatic text simplification: OneStopEnglish (Vajjala and Lučić, 2018). Each article has been rewritten by teachers to suit three levels of adult ESL learners (*elementary*, *intermediate*, and *advanced*) and are originally from the Guardian newspaper.[3] OneStopEnglishQA additionally contains high quality text comprehension questions that are explicitly linked to textual spans. In our study, we use only the *advanced* and *elementary* texts and ask the same questions per article independent of the level. Figure 3 illustrates an example of a comprehension question and the relevant textual spans for both an *advanced* and *elementary* paragraph. For each text, we present three comprehension questions that have been strategically chosen to necessitate scrolling. The texts presented to participants are chosen randomly, however a participant will never be shown the same article at differing levels.

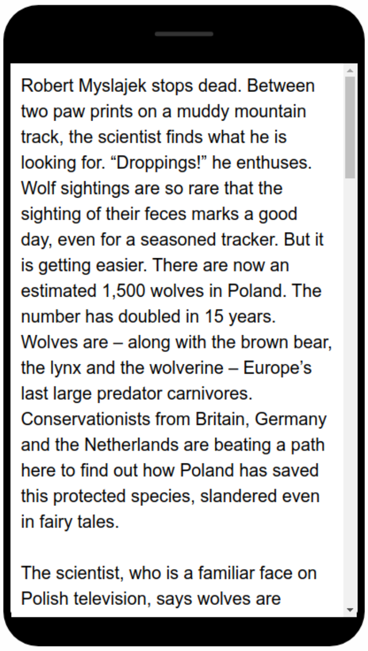
Task and Procedure

The task required participants to read articles and then answer three comprehension questions. Initially, participants were shown instructions as well as a placeholder text with questions before confirming that they were ready to begin. Participants were presented with two texts, one at an *advanced* level and one at an *elementary* level. Texts were shown one at a time in a random order. Participants were immediately given the comprehension questions after reading the text. The instructions stated that participants must read the article and then answer the comprehension questions. However, the text was still accessible during question answering and we record these interactions separately to reading. Additionally, participants were informed that they would be awarded a financial bonus on top of their base rate of pay for each question answered correctly. The decision to award a bonus was to encourage participants to read the text carefully. When presented with an article, participants were not able to progress until at least 90 seconds had elapsed to encourage engagement and prevent immediate skipping. Once the participant had read the two texts and answered all questions, a demographic questionnaire was presented.

Implementation

We used the Qualtrics platform to create the survey and added a custom front end implementation using HTML, CSS and JavaScript. The experiment interface is illustrated in Figure 2 and was displayed via a browser window. The article text was presented to participants in a restricted window of size 1080

[3] https://www.theguardian.com/uk



Robert Myslajek stops dead. Between two paw prints on a muddy mountain track, the scientist finds what he is looking for. “Droppings!” he enthuses. Wolf sightings are so rare that the sighting of their feces marks a good day, even for a seasoned tracker. But it is getting easier. There are now an estimated 1,500 wolves in Poland. The number has doubled in 15 years. Wolves are – along with the brown bear, the lynx and the wolverine – Europe’s last large predator carnivores. Conservationists from Britain, Germany and the Netherlands are beating a path here to find out how Poland has saved this protected species, slandered even in fairy tales.

The scientist, who is a familiar face on Polish television, says wolves are

Figure 2: Example of the reading interface presented to participants.

Elementary

The brown beer bottle was in the water for 101 years. A fisherman found it in the Baltic Sea off the Northern city of Kiel. Holger von Neuhoff a curator at the museum said this bottled message was the oldest he had ever seen.

Advanced

The brown beer bottle, which had been in the water for 101 years, was found in the catch of Konrad Fischer, a a fisherman, who had been out in the Baltic Sea off of the northern city of Kiel. Holger von Neuhoff, curator for ocean and science at the museum, said that this bottled message was the oldest he had come across.

Q: How was the bottle discovered?
a: A fisherman discovered it
b: It washed up on the shore of Kiel
c: Holger von Neuhoff discovered it
d: A boy discovered it while playing in the sea

Figure 3: Example paragraph from *elementary* and *advanced* texts. In the figure, answer spans are highlighted in green and distractor spans in pink.

by 1920 CSS px, a density-independent measure, based on the dimensions of an average Android device. The font used in the study was sans-serif size 18pt.

We log an event whenever the participant scrolls on the text. Events are logged every $100ms$ and include a timestamp indicating the elapsed time as well as the scroll y-axis offset in pixels (the vertical distance from the top of the text to the current location).

3.1 Preprocessing

We collected responses from 600 participants in total. However, the dataset was initially preprocessed to mark entries where the participant had not sufficiently engaged with reading the text. Given the size of the screen and text lengths, scrolling was necessary in order to read the text. We only included participants in our analysis if their scroll pattern indicated that they had attempted to read the article. If no scrolling had been logged, or the participant had not reached the half way point, the entry was not included in the analysis. Removing these entries resulted in 518 participants and 1036 articles marked as read.

Interaction Measures

As a preliminary analysis, a range of interaction metrics were extracted from the scrolling behaviour of participants. The computation of these follow the standard of those used in prior scroll research (Hinckley et al., 2002). These measures are outlined below.

ELAPSED TIME: The total reading time is recorded to produce a *read time* in seconds.

SPEED: The scroll speed (s) is calculated for each scroll interaction using $s = d/t$, where d represents the distance in pixels and t the time taken in ms. The average, minimum and maximum scroll speed are calculated.

ACCELERATION: Scrolling acceleration (a) is computed for each interaction. This is calculated with the following formula: $a = (v - u)/t$ where v represents the final scroll speed, u the initial speed and t denotes the time taken in ms. The average, minimum and maximum scroll acceleration are calculated.

TEXT REGRESSIONS: Scrolling typically takes place in a linear vertical fashion. Whilst reading, areas of text may require re-covering resulting in upward scrolling actions. We count the number of times the participant scrolled upwards to recover areas of text.

Norm. Measures		*Elementary*			*Advanced*			
$(/length)$	$\times 10^n$	$\bar{X}$	σ	r	$\bar{X}$	σ	r	$Sign.$
Read time (s)	$\times 10^2$	3.4	3.8	0.10	2.1	1.9	0.14 ●	$p < 10^{-18}$
Regression num	$\times 10^{-3}$	5.9	9.3	-0.09	7.1	8.0	-0.03	$p < 10^{-3}$
Min speed (px/ms)	$\times 10^{-5}$	4.1	11	-0.17 ●	3.3	8.7	0.01	-
Max speed (px/ms)	$\times 10^{-3}$	2.4	1.7	-0.23 ●	1.7	1.3	-0.12	$p < 10^{-12}$
Avg speed (px/ms)	$\times 10^{-4}$	7.9	6.0	-0.35 ●	6.1	4.6	-0.26 ●	$p < 10^{-14}$
Min acc (px/ms^2)	$\times 10^{-6}$	-5.9	8.5	0.12	-4.4	5.8	0.08	$p < 10^{-5}$
Max acc (px/ms^2)	$\times 10^{-6}$	-4.7	8.7	-0.11	-3.4	5.8	-0.08	$p < 10^{-3}$
Avg acc (px/ms^2)	$\times 10^{-7}$	-3.9	13	0.07	-2.0	6.7	0.07	$p < 10^{-3}$

●: $p < 0.01$ $\bar{X}$: Mean value σ: Standard deviation

Table 1: Interaction measures for 518 participants across *elementary* and *advanced* texts. Measures have been normalised according to text lengths. The correlation (r) of measures with comprehension scores is presented (p-values have been Bonferroni-corrected). The statistical significance of group differences (Sign.) is calculated using a mixed-effects model.

4 Results

Readability Measures

Table 1 shows the mean values ($\bar{X}$) and the standard deviation (σ) of reading interactions normalised by text length. We additionally present the correlation (r) of these with the comprehension scores of participants. P-values have been Bonferroni-corrected to account for multiple test conditions. The results show that three measures correlate significantly with participant scores on *elementary* texts and two for *advanced*. For the *elementary* articles, all speed measures correlate significantly with participant scores. The negative correlation shows that the slower the speeds, the higher the subsequent score. For advanced texts, the average scroll speed also negatively correlates with score. Whereas the time taken to read the article positively correlates with the subsequent score.

When considering the mean values across interactions, we see that all speed and acceleration measures are lower for *advanced* texts. This finding shows that, on average, the speed and acceleration of scrolling is slower on texts that are more difficult. Additionally, the number of regressions is larger for *advanced* texts than for *beginner*. Therefore, participants were more likely to recover areas of text when it is at a higher level. Finally, the standard deviation is larger for all measures on the *elementary* texts. This implies that there is more variance in reading interaction styles when the text is at a lower level.

We consider whether reading interactions differ significantly depending on the level of the text. We calculate significance using Satterthwaite's method (Kuznetsova et al., 2017), applied to a mixed-effects model that treats participants and texts as crossed random effects.[4] All measures are found to be statistically significant apart from the minimum speed. The most significant measure ($p < 10^{-18}$), is the normalised time a participant spent on the text. Counter-intuitively, the time taken, on average, is longer for the *elementary* texts than for *advanced*. However, when we consider proficiency groups, participants with lower proficiency levels took longer on *elementary* texts compared to *advanced*. This suggests that for readers at a low level the text is too difficult to engage with, resulting in the skipping of content. All acceleration measures, the maximum and average speed as well as the number of regressions differ significantly depending on whether text is *advanced* or *elementary*. These findings support the case that there are different reading interactions for texts depending on their complexity.

Predicting Readability

We perform experiments to investigate whether scroll-based reading interactions can be used to predict the level of a given text. As the texts in our dataset have been read by multiple participants, we produce features by taking the average of the statistically significant scroll measures (as displayed under Sign. in Table 1). Prior research has shown that simple methods of combining group interactions, such as averaging or majority voting, can be

[4] Using R formula notation, the model is: $measure \sim readability + (readability|participant) + (readability|text)$. Tests were performed using the lme4 and lmerTest R packages by Bates et al. (2014).

System	*N*	*Precision*	*Recall*	*F-Score*
Scroll only	12	0.80	0.78	0.77
VAJJALA-2018	155	0.88	0.85	0.84
VAJJALA-2018 + *Scroll*	160	0.92	0.88	0.88
Baseline (length + LR)	6	0.91	0.87	0.88
Baseline + *Traditional*	15	0.92	0.90	0.89
Baseline + *Discourse*	24	0.81	0.82	0.80
Baseline + *Syntactic*	24	0.87	0.85	0.84
Baseline + *Psycholinguistic*	12	0.87	0.88	0.87
Baseline + *Scroll*	18	**0.98**	**0.97**	**0.96**

Table 2: Table showing the total number of features per system (N) as well as the precision, recall and f-score of models trained using 10-fold cross-validation on the OneStopQA dataset.

Scroll features	*F-Score*
All (¬Norm + Norm)	**0.77**
¬ Normalised	0.64
Normalised	0.61

Table 3: Readability prediction using aggregate interaction features.

highly robust (Genre et al., 2013; Clemen, 1989; Ertekin et al., 2012). We produce two sets of scroll features using the number of regressions, the max and average scroll speed and the max, min and average scroll acceleration. One set is normalised by text length (Normalised) and the other not (¬ Normalised). We then train a support vector machine (SVM) to predict whether a text is *advanced* or *elementary* using these features. This classifier was chosen as it has been shown to consistently yield better results compared to other statistical models when predicting text readability (Kate et al., 2010). All reported results are obtained using stratified 10-fold cross-validation.

Table 3 presents the f-score of the resulting models. To the best of our knowledge, this is the first work aiming to predict readability using scroll-based features. The best result (0.77) is achieved when using both sets of scroll features. The ¬ Normalised feature set performs better than Normalised. This is likely due to the fact that these features contain signal on the length of texts.

Interestingly, filtering interactions by specific sub-groups can produce better scores. The best results are achieved, using both sets of scroll measures, when we only include the interactions of participants aged 25-34 – resulting in an f-score of 0.81. There are known differences in computer interaction styles based on age (Schneider et al., 2008). It therefore follows that aggregate measures from an audience with a more consistent interaction style would result in better features.

The use of scroll features alone produces an f-score of 0.77. We compare our result to a state-of-the-art readability system by Vajjala and Lučić (2018) referred to as VAJJALA-2018. This system uses a multilayer perceptron classifier and has been shown to outperform BERT-based approaches on the OneStopEnglish dataset (Martinc et al., 2019). The system relies on 155 hand-crafted features which are grouped into six categories: traditional metrics, word features, psycholinguistic, lexical richness, syntactic and discourse features. The VAJJALA-2018 system outperforms the use of scroll features alone when trained to predict if texts are *advanced* or *elementary* on our dataset. However, we are able to improve the f-score of this state-of-the-art system by 4% when adding aggregate scroll features. This is an interesting result as it shows that 1) these features are highly complementary to existing readability systems and 2) these features represent an aspect of textual complexity not covered by the current 155 features. This may be due to scroll-based features capturing a notion of conceptual complexity which plays a vital role in text understanding and maintaining a reader's interest (Štajner and Hulpuș, 2020).

We also investigate how scroll-interaction features compare to classical readability feature sets (an overview of feature sets is included in Appendix A). To do this, we initially create a highly competitive baseline by training an SVM using the length of text and measures of lexical richness. Due to the nature of the OneStopEnglish dataset, length is an extremely informative feature. This is because, on average, the word length for *advanced* texts (915) is almost always higher than for the simplified *elementary* texts (599). Vocabulary knowledge, including lexical diversity and richness, are principal components in reading comprehension (Collins-Thompson, 2014). The importance of lexical richness has been investigated from the perspective of second language acquisition (Lu, 2012), as well as in readability systems (Vajjala and Meurers, 2012; Vajjala and Lučić, 2018; Xia, 2019). We opted to use features in our baseline system that were highly informative but would not require extensive text processing. A key advantage of lexical richness measures is that they are a function of the number of word types (T) and the total text length (N). Therefore, they can be calculated quickly and are robust to noisy or broken text. The specific measures of lexical richness we include are type-token ratio (TTR), which is the ratio of the number of unique word tokens to the total number of word tokens in a text, Root TTR ($T/\sqrt{N}$), Corrected TTR ($T/\sqrt{2N}$), Bilogarithmic TTR ($LogT/LogN$) and Uber Index ($Log^2T/Log(N/T)$). The resulting baseline approach outperforms the VAJJALA-2018 system. However, it should be noted that the VAJJALA-2018 system was originally designed to differentiate between *elementary*, *intermediate* and *advanced* texts which is a more nuanced and difficult task.

By performing ablation tests, we see how scroll-interaction features compare to classical sets of readability features when combined with a strong baseline system (Length + LR). We find that adding sets of discourse, syntactic or psycholinguistic features degrades classifier performance. The addition of traditional measures, such as Flesch-Kincaid score, produces a small improvement. However, adding aggregate scroll features produces a marked improvement on the baseline resulting in an f-score of 0.96. The addition of scroll-interaction features improves performance beyond all other classical feature combinations. A key advantage to this approach is that the result is achieved with only 18 features which are highly robust to noisy text.

Subjective Readability

Readability approaches typically produce an objective numerical score which often corresponds to a suggested level (Fry, 2002). In our previous experiments, we predict the readability of text as defined by such preordained levels (i.e., *elementary* or *advanced*). However, the readability of text can also be defined in a more subjective and idiosyncratic manner. Such techniques have been referred to as *levelling*, and are similar to readability in that they determine difficulty but are subjective (Clay, 1991). Levelling integrates a reader's background and experience with objective readability by understanding what contributes to readability for differing audiences. In this section, we investigate what reading-interactions could tell us about an individual's text comprehension or L1.

In our study, we record comprehension scores to evaluate the understanding and readability of text for individuals. Participants are asked three questions per article assessing the reader's understanding of what they have read. Figure 4 shows the average score of participants grouped by their self-reported proficiency. We see that the average score decreases in line with the reported proficiency.

In the fashion of levelling, we investigate the correlation of features with the comprehension scores of audiences. We observe whether the feature correlations, at both the text and interaction level, vary depending on the first language (L1) of the group. We focus on English (n = 350) and Tamil (n = 101). Tamil was chosen as it was the second most represented L1 after English.

We find there is a statistically significant association between the average scroll speed and scores of participants for both Tamil and English ($p < 10^{-4}$). The correlation is negative, indicating that the faster the average speed of reading, the lower the subsequent score. When considering the English L1 audience, the only statistically significant correlation with score was the average reading speed. However, for Tamil readers there were three statistically significant correlates with ($p < 10^{-4}$), in order of strength these were: the mean age of acquisition for vocabulary in the text (AoA) as reported by Kuperman et al. (2012); the average reading speed and the mean AoA of word lemmas. The negative correlation shows that the higher the AoA of vocabulary, the lower the score for this group. This finding supports prior work emphasising the importance of AoA as a factor when simplifying texts for second language learners (Crossley et al., 2007).

The average scrolling speed correlates significantly with scores for both English and Tamil L1 audiences. We investigate whether the distribution of average reading speeds differs based on the reader group. Figure 5 presents a histogram of the average reading speeds. In both groups, we observe

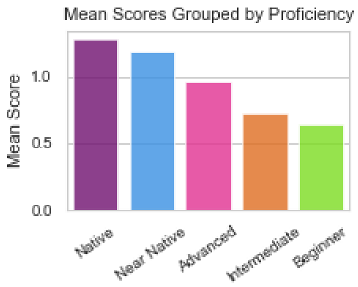
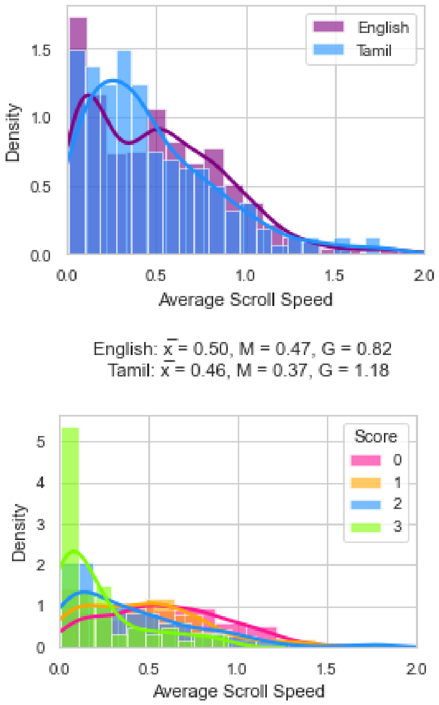


Figure 4: Bar chart showing the average scores (out of three) for texts across proficiency levels.

a positively right skewed distribution, as speed is a vector quantity and a range boundary occurs at 0. When comparing speed distributions, the mean and median speeds are higher for English (0.50, 0.47) than Tamil (0.46, 0.37). The higher median and mean measures for English interactions are likely due to a higher familiarity yielding a faster reading speed. The L1 English group additionally achieve a higher score on average (1.18) than the Tamil audience (0.98), despite the faster scroll speed.

In Figure 5, we consider how the distribution of average reading speeds varies for participants according to the score they attained. The higher the score, on average, the lower the scroll (reading) speed. For the lower scores, we see that the distribution has a wider spread. According to the self-reported proficiency ratings, the *advanced* texts would likely have been too difficult for a proportion of participants. A higher scrolling speed indicates that a reader is skipping content without properly reading, perhaps due to the level being too high for the reader to sufficiently engage with. Finally, there is a statistically significant positive correlation between the average reading speed of a participant and their reported proficiency ($p < 10^{-4}$), further supporting the notion that reading interactions vary based on ability.

Being able to understand how on-device reading interactions vary according to an individuals' L1 and comprehension is incredibly useful. Such information could be used in text simplification systems, or in a 'levelling' manner to match the appropriate level of text to a given reader. Such applications are especially useful for individuals who are learning a language.



English: $\bar{x}$ = 0.50, M = 0.47, G = 0.82
Tamil: $\bar{x}$ = 0.46, M = 0.37, G = 1.18

Figure 5: Histograms showing the density of average scroll speed for Tamil and English as well as across participant scores.

5 Conclusions

To conclude, the use of scroll features for judging readability has numerous benefits. Such measures are language agnostic, unobtrusive and are robust to noisy text. Furthermore, implicit user feedback allows an insight into readability at an individual level, thereby allowing for a more inclusive and personalisable assessment. We present a 518 participant study to investigate the impact of text readability on reading interactions. In this paper, we confirm that there are statistically significant differences in the way that readers interact with *advanced* and *elementary* texts, and that the comprehension scores of individuals correlate with specific measures of scrolling interaction. We demonstrate that, even with a simple model, aggregate scroll interactions can be used to predict readability. Moreover, we show that individual scroll behaviour can provide an insight into the subjective readability for an individual. Finally, we improve a state-of-the-art readability classifier with the integration of scroll-interaction features, demonstrating that interaction features are highly complementary to traditional linguistic approaches. In future work, we will focus on investigating which aspects of readability scroll-based measures index.

References

Daniel J Acheson, Justine B Wells, and Maryellen C MacDonald. 2008. New and updated tests of print exposure and reading abilities in college students. *Behavior research methods*, 40(1):278–289.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. Starc: Structured annotations for reading comprehension. *arXiv preprint arXiv:2004.14797*.

Jeanne Sternlicht Chall. 1958. *Readability: An appraisal of research and application*. 34. Ohio State University.

Robert Chen, Roger Levy, and Tiwalayo Eisape. 2021. On factors influencing typing time: Insights from a viral online typing game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Marie M Clay. 1991. *Becoming literate: The construction of inner control*. Heinemann Educational Books.

Mark Claypool, Phong Le, Makoto Wased, and David Brown. 2001. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40.

Robert T Clemen. 1989. Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, 5(4):559–583.

Kevyn Collins-Thompson. 2013. Enriching the web by modeling reading difficulty. In *ESAIR*, pages 3–4.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 193–200.

Scott A Crossley, Philip M McCarthy, and Danielle S McNamara. 2007. Discriminating between second language learning text-types. In *FLAIRS Conference*, pages 205–210.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Seyda Ertekin, Haym Hirsh, and Cynthia Rudin. 2012. Learning to predict the wisdom of crowds. *arXiv preprint arXiv:1204.3611*.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57.

Edward Fry. 2002. Readability versus leveling. *The Reading Teacher*, 56(3):286–291.

Véronique Genre, Geoff Kenny, Aidan Meyler, and Allan Timmermann. 2013. Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

Gene Golovchinsky, Morgan N Price, and Bill N Schilit. 1999. From reading to retrieval: freeform ink annotations as queries. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

Ana V González-Garduno and Anders Søgaard. 2018. Learning to predict readability using eye-movement data from natives and learners. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association* *for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79.

Ken Hinckley, Edward Cutrell, Steve Bathiche, and Tim Muss. 2002. Quantitative analysis of scrolling techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 65–72.

Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, volume 37, pages 18–28. ACM New York, NY, USA.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

George Roger Klare et al. 1963. Measurement of readability.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Alexandra Kuznetsova, Per B Brockhoff, Rune HB Christensen, et al. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13):1–26.

Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katharina Reinecke. 2019. The impact of web browser reader views on reading speed and user experience. CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2019. Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779*.

Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. 2018. Understanding the needs of searchers with dyslexia. CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Nicole Schneider, Janet Wilkes, Morten Grandt, and Christopher M Schlick. 2008. Investigation of input devices for the age-differentiated design of human-computer interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 144–148. SAGE Publications Sage CA: Los Angeles, CA.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

Sanja Štajner and Ioana Hulpuș. 2020. When shallow is good enough: Automatic assessment of conceptual text complexity using shallow semantic features. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1414–1422.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Ryen W White, Ian Ruthven, and Joemon M Jose. 2002. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 57–64.

Menglin Xia. 2019. *Text readability and summarisation for non-native reading comprehension*. Ph.D. thesis, University of Cambridge.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Beverly L Zakaluk and S Jay Samuels. 1988. *Readability: Its Past, Present, and Future.* ERIC.

Yixin Zou, Shawn Danino, Kaiwen Sun, and Florian Schaub. 2019. You 'might' be affected: An empirical analysis of readability and usability issues in data breach notifications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

6 Appendix

A - Traditional Readability Feature Sets

The readability feature sets for the OneStopEnglish dataset were provided by (Vajjala and Lučić., 2018)[5]. Please refer to their work for further details. An overview of features is provided:

Traditional

- Avg. number of characters per word
- Avg. number of syllables per word
- Avg. sentence length
- Flesch-Kincaid score
- Coleman-Liau readability formula
- SMOG grade

Discourse

- Word overlap features - content word, noun, stem and argument overlap at local (between adjacent sentences) and global (between any two sentences in a text) levels.
- Entity transitions - features based on the transitions between the syntactic roles of entities.
- Co-reference chains - features based on noun phrases, nouns and pronouns and determiner usage.
- Num. of referential expressions.
- Num. of discourse and non-discourse connectives and all connectives

Syntactic

- Num. NPs per sentence (NumNP)
- Num. VPs per sentence (NumVP)
- Num. PPs per sentence (NumPP))
- Avg. length of a NP (NPSize)
- Num. Dependent Clauses per sentence (NumDC)
- Num. Complex-T units per sentence (NumCT)
- Num. Co-ordinate Phrases per sentence (Co-Ord)
- Num. SBARs per sentence (NumSBAR)
- Avg. Parse Tree Height (TreeHeight)
- Avg. length of a VP (VPSize)
- Avg. length of a PP (PPSize)
- Mean length of clause (MLC)
- Mean length of a sentence (MLS)
- Mean length of T-unit (MLT)
- Num. of Clauses per Sentence (C/S)
- Num. of T-Units per sentence (T/S)
- Num. of Clauses per T-unit (C/T)
- Num. of Complex-T-Units per T-unit (CT/T)
- Dependent Clause to Clause Ratio (DC/C)
- Dependent Clause to T-unit Ratio (DC/T)
- Co-ordinate Phrases per Clause (CP/C)
- Co-ordinate Phrases per T-unit (CP/T)
- Complex Nominals per Clause (CN/C) – Complex Nominals per T-unit (CN/T)
- Verb phrases per T-unit (VP/T)

Psycholinguistic

Norms from MRC which were compiled by Gilhooly and Logie (1980) for 1903 unique words including:

- Avg. word age of acquisition
- Avg. word Familiarity
- Avg. word concreteness
- Avg. word imagability
- Avg. word meaningfulness

[5] https://zenodo.org/record/1219041