Log-based Reading Speed Prediction: a Case Study on *War and Peace*

Igor Tukh[1], Pavel Braslavski[1,2,3], and Kseniya Buraya[4]

[1] Higher School of Economics, Saint Petersburg, Russia
[2] Ural Federal University, Yekaterinburg, Russia
[3] JetBrains Research, Saint Petersburg, Russia
[4] ITMO University, Saint Petersburg, Russia
{igor-tukh,pbras}@yandex.ru, ks.buraya@gmail.com

Abstract. In this exploratory study, we analyze reading behavior using logs from an ebook reading app. The logs contain users' page turns along with time stamps and page sizes in characters. We focus on 17 readers of *War and Peace* by Leo Tolstoy, who read at least 80% of the novel. We aim at learning a regression model for reading speed based on shallow textual (e.g. word and sentence lengths) and contextual (e.g. time of the day and position in the book) features. Contextual features outperform textual ones and allow to predict reading speed with moderate quality. We share insights about the results and outline directions for future research. The analysis of reading behavior can be beneficial for school education, reading promotion, book recommendation, as well as for traditional creative writing and interactive fiction design.

Keywords: reading speed, text difficulty, reading behavior, user modeling

1 Introduction

Reading speed is an indicator that is often used to assess reading skills in one's mother tongue or in a foreign language. Reading speed is commonly associated with cognitive abilities and personal effectiveness. There are numerous speed reading techniques aimed at increasing the speed of reading by several times. Reading speed is an indirect indicator of readers' interest in what they read. Most of the experiments and practical tests to measure reading speed are conducted in controlled settings, usually using short standardized educational or technical texts. One can assume that fiction reading is a more complex phenomenon, and reading speed of fiction works is affected by a variety of factors – besides superficial text complexity, plot, suspense, style, and reader's engagement may come into play. For example, an Argentine Canadian essayist and librarian Alberto Manguel wrote about different ways he used to read in his adolescence [11]: "First, by following breathlessly, the events and the characters without stopping to notice the details, the quickening pace of reading sometimes hurtling the story beyond the last page... Secondly, by careful exploration, scrutinizing the text to understand its ravelled meaning, finding pleasure merely in the sound of the words or in the clues which the words did not wish to reveal, or in what I suspected was hidden deep in the story itself, something too terrible or too marvellous to be looked at." The former way of reading Manguel attributed to adventure fiction including *Odyssey*, while the latter – to works by Lewis Carrol, Dante, Kipling, and Borges. Despite many evidence like this, experimental investigation of fiction reading and analysis of reading 'in the wild' are few. This is partly because fiction reading remains mostly an intimate process avoiding external intrusion. Nowadays, ebooks reading logs can provide rich and fine-grained information about reading behavior.

In this study, we examine the reading speed of a novel by several readers over a considerable period of time. To this end, we use a log of a mobile reading application. In this exploratory study, we focus on 17 users, who read a significant portion (80% or above) of *War and Peace* by Leo Tolstoy. The log contains timestamps of users' turning pages on the mobile device and the sizes of these pages in characters. We align these log records with the text of the novel and split it into corresponding pages. For each reader we build a model that predicts their reading speed of a page based on its content or/and context and test the model on held-out data. As content features we employ various shallow features traditionally used in readability formulas such as word/sentences length and rare word ratio, whereas context features include time of the day, day of the week, and position in the book. The difficulty of reading speed prediction based on log data lies in the fact that, in contrast to laboratory settings, we can't control external conditions and know practically nothing about the readers whose behavior we study. Further obstacle is a non-perfect alignment of text and log entries. In addition, the range of speeds that we observe in the data is very wide, and corresponds not only to actual reading, but also to idle periods, skimming, and flipping-through the book.

The results show that we are able to obtain predictions of moderate quality if we narrow the range of considered speed values. Contextual features provide better predictions than textual ones. In future studies, we plan to expand the set of book titles and the population of readers, enrich the feature set, as well as address the problem of distinguishing actual reading from idling, skimming, and fast-forward browsing.

Analysis of real-life reading behavior can be of interest for school education, reading promotion, book recommendation, as well as for creative writing. In case we are able to connect reading speed to reader engagement, it can be used in interactive fiction, where a book's plot can change dynamically depending on the reader's explicit or implicit feedback.

2 Related Work

Standardized reading speed tests play an important role in educational practice and cognitive ability testing. An example of such a test is International Reading Speed Texts (IReST) [17]. The authors develop a set of comparable texts in 17 languages. The authors use educational text intended primarily for school students. Based on the experiments involving 25 adult participants, they report average reading aloud speed and standard deviation for Russian – 986 (175) characters/min (white spaces and punctuation are not counted in). Another experiment on silent reading speed in Russian [5] involves 533 adult participants and uses a text on a philosophical topic. The authors report much higher average speed and surprisingly lower standard deviation – 1,596 (49) characters/min (it's not clear from the paper how white spaces and punctuation are treated). Eye tracking has been used for studying cognitive and psychological aspects of reading for over a century.

Early work focused mainly on the interpretation of eye movements and their relationship to cognitive processes of language processing and text comprehension. Different characteristics of the text (e.g., complexity and typographic values), specific language phenomena (different kinds of linguistic ambiguity, anaphora, semantic relationships between words, etc.), as well as individual characteristics of the reader (background knowledge, reading proficiency, etc.) were studied using eye tracking, see a comprehensive survey [15].

Kunze et al. [8] proposed using eye tracking for logging individual reading behavior analogously to fitness tracking or food logging mobile applications. The approach uses built-in cameras of the mobile devices for eye-tracking and delivers a summary of volume, speed and schedule of reading. Unfortunately, no results of such tracking studies have been reported yet.

A relevant task in the context of our study is eye tracking-based classifier into reading/skimming behavior [1]. In the skimming mode readers digest the text at speed up to two times higher than normal, at cost of missing details. Masson experimented with reading speeds beyond normal reading (200–300 words/min for English) [12]. He found out that at the speed of 375 words/min the reader can still comprehend main ideas expressed in the text, but at the speed of 600 words/min (about 3,000 chars/min) the reader is unable to follow even the main topics in the text. Such 'speed reading' can be still helpful in spotting concrete information in the text.

Nell [13] investigated reading speed in pleasure, or *ludic* reading. Within 30-minute reading sessions in the lab, he found a high degree of reading rate variability: average high/low per page speed ratio was 2.63 among 33 participants. The study also found out that readers slowed down on pages they liked most. However, since only page transitions were recorded within the experiment, it is not clear, whether this variability in due to slower 'linear' reading, or because most-liked passages were reread more than once. László and Cupchik [10] introduced two types of literary narratives – *action*- and *experience*-oriented – and measured reading speed, as well as subjective time perception during the reading of both. They showed that *action* fragments were read faster than *experience* fragments. The readers in the experiment associated *action* stories with lower reading difficulty and lower predictability. Brouwer et al. [4] used such physiological signals as EEG, ECG, skin conductance, heart rate, and respiration of readers to be able to distinguish emotionally intense vs. neutral sections of a novel.

Readability scores has been widely used to evaluate educational and instructional materials since 1920s. Traditional readability formulas are results of regression analysis that sees experimentally obtained text complexity as dependent variable and several computable text features as independent variables. Text complexity is measured for instance by reading time (normalized by the individual reading proficiency), post-reading questionnaires assessing text comprehension, or cloze tests. Text features used in readability formulas include various word lists (such as "easy", "hard", "abstract", "most frequent", etc. words), word and sentence length, number of prepositional phrases, etc.; see [7] for a comprehensive survey. In our study, we rely on these simple text difficulty signals. A more recent study [14] incorporates lexical, syntactic, and discourse features to predict text readability. We will consider more complex features for reading speed prediction in our future research.

There are some studies on news reading behavior on the Web. Constantinides et al. [6] use reading speed as a feature for news personalization. They distinguish three styles of news reading: detailed reading (up to 230 words/min), normal reading and skimming (230–700 words/min), and scanning (above 700 words/min). Lagun and Lalmas [9] analyze online news reading on sub-document level, which makes the study close to ours. The authors utilize viewport data and model user engagement with the content at sub-document level based on text features, as well as mine different news article reading patterns.

3 Data

In this study we use reading logs from Bookmate[5], a Russian mobile reading app. Upon installing the application, users get instant access to a free collection of several thousands of titles (mostly Russian classical novels); further they can choose from two subscription levels. Standard subscription grants a user access to the entire Russian book collection, excluding new arrivals, bestsellers, and business books. Premium subscription provides unlimited access to the entire collection. App logs used in the study correspond to almost 10 months – from January to October 2015. The data includes information about the users, books, and readings sessions. Detailed description of the dataset and reading behavior characteristics derived from the log can be found in [3].

Book data contains book ID, author, title and length in character.

User information includes their subscription type (paying/non-paying); some users indicated their gender and year of birth.

The main contents of the dataset are page turns.

Every record in the log contains the user and book IDs, the time stamp of a page turn, and the character range that corresponds to the turned page size, as well as additional auxiliary information. The size of the page in characters depends mainly on the screen size of the user's device.

In this exploratory study we opted for focusing on a single book – *War and Peace* by Leo Tolstoy (*W&P*). The novel is considered as one of the Tolstoy's

[5] https://www.bookmate.com

masterpieces and depicts life of several Russian noble families in the time of Napoleonic wars. The four-volume 1,300-page novel was first published in 1869. *War and Peace* is included in the Russian high-school curriculum; there are several film and television adaptations of the novel. It's interesting to note in the context of our study that there is an anecdotal evidence that female and male high-school students read the novel in different ways: girls skip battle scenes, while boys – balls and long dialogues.[6] *W&P* is also actively researched within the digital humanities, see for example [2].

For this study, we considered only readers, who have read at least 80% of the book, with earliest reading sessions within first 10% of the book.

We also removed users with the coverage above 120% and ended up with 17 readers.[7]

Then, we cleaned the data as follows.

We removed duplicate sessions, i.e. nearly simultaneous (up to five seconds apart from each other) sessions spanning the same text interval.

Readers don't read the novel linearly – we can observe backward and forward leaps in the data. Figure 1 illustrates reading sequence by one of the readers in the study: spikes below/above diagonal reflect backward/forward movements, respectively. Further, we removed distant 'backward jumps' within the book. We assume these records reflect navigational browsing, device orientation changing, waking up device, and similar behavior, thus contain no useful information in the context of the current study. We also removed sessions corresponding to the fragments shorter than 20 words (either headings or last lines of a chapter) that don't allow to calculate reliable content statistics.

On the next step, we calculate start/end positions for each page based on log data and mapped the intervals to the ebook file. This alignment can't produce perfect mapping with what the users saw on their screens and is a source of additional noise in the data. Based on two subsequent timestamps we calculate time the reader spent reading this page, except for the first page in the sequence. We set the reading time of the first page in the sequence to the session's average. Essentially, our final data consists of character ranges from *W&P* and such time intervals .

Table 1 describes main characteristics of the *W&P* readership in the study. It can be seen that the users vary greatly in terms of books accessed and time spent at reading in the app. Inspection of titles read by users with few books (e.g. #3, 7, 13, 15–17) suggests that these are high-school students reading free books from the list of assigned readings. This fact may influence results, since 'obligatory' reading behavior may differ from spontaneous leisure reading. Figure 2 shows overall distribution of reading speed for *W&P*, while Table 1 cites individual averaged speed values. It can be seen from the table that *unfiltered* average reading speed calculated upon log data is often significantly higher than normal

[6] We were unable to test this hypothesis due to incomplete data.

[7] To calculate coverage we summed up all character ranges in the log entries for a particular reader. Coverage above 100% occurs, when the same text spans are read or just flipped through several times.

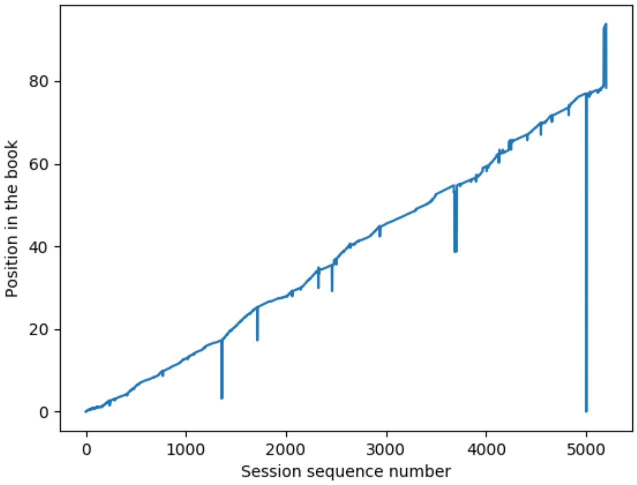

Fig. 1: Sequence of reading positions for reader #1

speed (1,000–1,500 characters/min, see above). It suggests that user often skim or just flip through the book.

To visualize reading behavior of the 17 readers in the study, we produced a heatmap of reading speeds, see Figure 3. To this end we divided the entire text of the *W&P* novel into 400 fragments of equal size (each fragment corresponds to about three pages of a printed book; we didn't account for volume and chapter division). Further, we normalized individual reading speeds by mapping them to [0..1] interval based on individual readers' min/max values and calculating new values for the fragments. Finally, we tried to rank the readers in such a way that readers with similar behavior are placed close to each other. As the figure shows, there is no common pattern in reading speeds. However, we can see that there are different types of readers: some of them read the novel with fewer speed variations, the other switch between slower and more accelerated reading.

Currently we have no reliable method to distinguish between idle periods (periods of user's inactivity followed by a page turn that appear as slow reading), normal reading, skimming, and fast-forward flipping. We set low/high thresholds for reading speed values 800/3,000 characters/min based on data reported in the literature and discard records outside this range. We assume that this span includes normal reading and skimming behavior. We investigate also the impact of threshold values on overall prediction quality (see below).

Table 1: 17 readers in the study: *$* – subscription type; *gender* as indicated in the user's profile; *#books* – total different books accessed and *total time* in hours spent on the service during the log's period; unfiltered *average reading speed* for *War and Peace* and estimated *time to complete* the novel in hours when reading at this pace; total *#sessions* in the log and *#filtered* sessions; *?* indicates missing data.

UserID	*$*	*gender*	*#books*	*total time*	*avg. speed*	*time to complete*	*#sessions*	*#filtered*
1	premium	m	48	124.8	3,040.92	19.85	5,202	3,387
2	premium	m	70	171.4	2,781.33	21.70	2,303	1,425
3	?	?	5	65.1	2,234.29	27.01	5,158	2,525
4	free	m	36	190.6	2,415.27	24.99	3,261	1,901
5	free	m	28	89.2	3,824.17	15.78	1,616	937
6	standard	?	26	213.9	1,272.85	47.42	6,171	4,490
7	free	?	4	68.3	3,619.21	16.68	2,936	1,757
8	free	?	19	100.4	1,224.99	49.27	7,623	5,601
9	standard	m	17	78.6	1,282.31	47.07	6,862	4,484
10	free	?	17	70.7	4,091.45	14.75	2,324	1,247
11	free	f	10	64.4	7,948.70	7.59	1,784	497
12	premium	?	17	67.3	2,300.29	26.24	4,736	3,620
13	free	?	3	63.4	3,965.01	15.22	2,576	1,223
14	free	?	10	44.0	1,895.32	31.84	5,004	3,891
15	free	m	4	38.2	3,126.40	19.30	2,024	1,681
16	free	f	6	52.8	2,222.29	27.16	3,837	3,054
17	?	?	4	30.0	2,071.25	29.14	4,105	3,219

4 Reading Speed Prediction

So far, we have text fragments (pages), their lengths in characters, and time the app user spent reading this fragment (two latter parameters deliver reading speed). We represent each text fragment as a vector of features, split the data corresponding to each reader into train and test subsets, learn a regression model, and evaluate it.

We implement two groups of features:

1. Text features:
   - *#Words* – number of words on the page;
   - *#Sentences* – number of sentences on the page, incomplete sentences are still counted in;
   - *Average word length in characters*;
   - *Average sentence length in characters*;
   - *#Rare words* based on unigram frequencies in the Russian National Corpus[8]
   - *#Finite verbs* and *#Nouns* based on *mystem* POS tagger.[9]
2. Context features:
   - *Hour of the day*: 0..23;
   - *Is_weekend*: 0 if weekday; 1 otherwise;

[8] http://ruscorpora.ru/corpora-freq.html

[9] https://tech.yandex.ru/mystem/

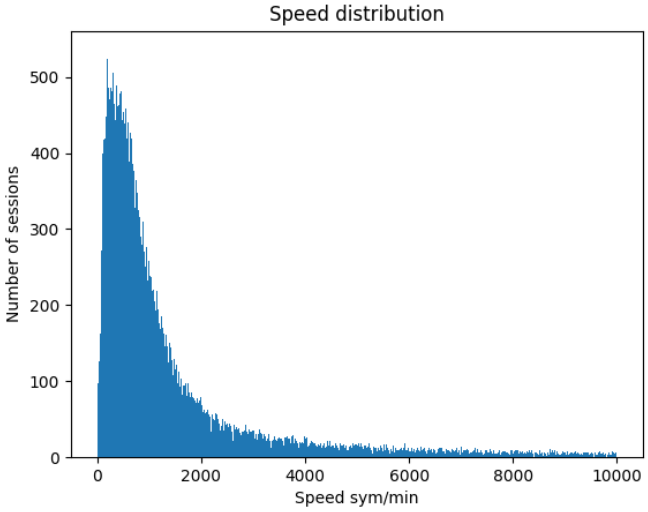

Fig. 2: Distribution of *W&P* reading speed

- *Position* of reading in percentage of the whole book.

We perform reading speed prediction based on text and context features separately, as well as on their combination.

We make predictions for each user independently. The data is split into train/test sets in the ratio 80/20 using two approaches:

- *Ordered*: sessions for a user are split in chronological order;
- *Random*: sessions for a user are split randomly.

We experiment with several regression methods implemented in *scikit-learn* library[10] and opted for Lasso [16] that delivered best results. We use Mean Absolute Error (MAE) as loss function. We also investigate the impact of different upper and lower speed thresholds on overall performance. We ran experiments for the lower threshold in the range from 100 to 800 characters/min and for the upper threshold in the range from 2,500 to 6,000 characters/min.

5 Results and Discussion

Table 2 summarizes the results of reading speed predictions for individual readers using different sets of features and data splits. As can be seen from the table, the quality of prediction differs significantly from reader to reader. One has to keep in mind that speed distributions for individual readers also vary greatly (see

[10] `https://scikit-learn.org/`

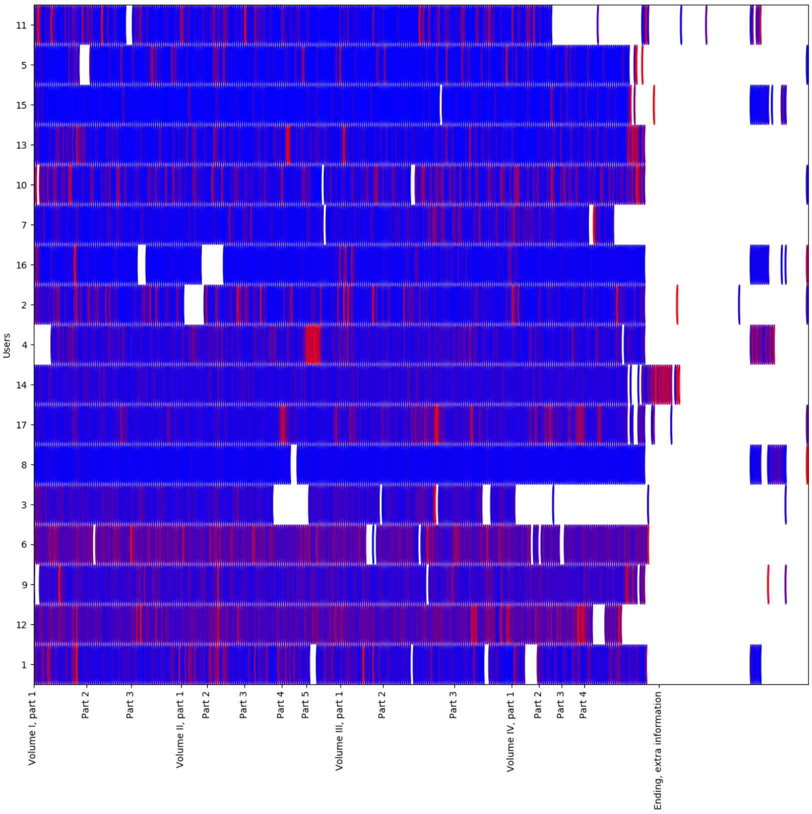

Fig. 3: Normalized reading speeds by 17 readers across *W&P*. The book is divided into 400 fragments of equal size; individual speed values are mapped to [0..1] interval; lower/higher speeds correspond to cold/hot parts of the spectrum, respectively; blanks correspond to missing data. Vertical ranking is aimed at placing readers with similar behavior close to each other.

Table 1, Figure 3). In most cases, speed predictions based on simple contextual features are more accurate. In some cases, combination of textual and contextual features reduces the error (cf. readers #3, 15–17). However, these results are only marginally better than predictions based solely on contextual features. In the case when textual features provide a better prediction compared to contextual ones (#10), the gain is neglectable. The impact of different data splits is mixed: for some readers, the error is lower on the ordered split, for the others – on the random split. This probably reflects different reading styles: for some readers, the behavior is consistent throughout the book, for the others it changes from the beginning of the novel to the end. It should be noted that in the case of contextual/combined features and random split the position within the book is still taken into account through the corresponding feature.

Table 2: Mean absolute error (MAE) for reading speed predictions in characters/min for different feature sets (*text* only, *context* only, and *combined* text & context) and train/test splits (*ordered*/*random*). Low/high speed thresholds are fixed at 800/3000 char/min. Best values for each reader are in **bold**; best results based on *context* features are underlined.

UserID	*Text*		*Context*		*Combined*	
	Ordered	*Random*	*Ordered*	*Random*	*Ordered*	*Random*
1	487.08	480.36	**442.92**	451.41	491.76	443.42
2	245.73	213.71	**159.68**	173.45	286.81	245.47
3	464.85	479.98	452.41	519.26	**445.17**	497.89
4	449.22	742.82	**272.28**	273.31	441.30	546.80
5	183.55	190.31	**170.41**	238.93	230.28	195.72
6	213.09	247.29	**160.41**	169.20	242.75	301.02
7	523.77	467.71	556.72	**438.90**	477.06	557.49
8	319.76	337.85	242.19	**223.53**	304.79	281.13
9	553.60	511.86	**253.20**	288.58	550.86	572.51
10	**365.00**	741.51	394.38	365.47	460.32	513.33
11	1262.14	1347.58	**499.30**	524.66	1472.18	1894.76
12	581.26	809.61	374.96	**330.70**	474.70	756.10
13	589.69	412.75	**351.20**	352.27	613.75	859.96
14	664.30	307.85	297.49	**293.65**	837.69	295.41
15	245.60	220.81	240.76	223.66	230.00	**218.22**
16	395.09	291.42	266.35	261.30	265.07	**255.40**
17	312.45	318.45	424.45	303.71	438.03	**303.06**

We investigated the effect of cut-off thresholds on speed prediction; the results for the lower and upper thresholds are presented in Figure 4. For the majority of readers, varying lower threshold from 100 to 800 characters/min has almost no effect. We can assume that their reading behavior doesn't change within this range. For four readers (#4, 12, 13, 17) the error drops with increasing the lower threshold. Supposedly, by increasing cut-off value we eliminate idle sessions that are different from actual reading. Results for the reader #11 demonstrate opposite behavior. These results suggest these differences must be taken into account when modeling reading behavior. We plan to tackle this problem in our future research.

6 Conclusions and Future Work

We conducted an exploratory study aimed at predicting reading speed of *War and Peace* fragments based on the log of an ebook application. We used two sets of features – textual and contextual. Somewhat surprisingly, simple contextual features significantly outperformed predictions based on textual features traditionally used in readability formulas.

In the future, we plan to use a larger collection of books and a larger population of readers. We will perform a more thorough data cleansing and segmentation of the user base by their behavior. We will address a more accurate

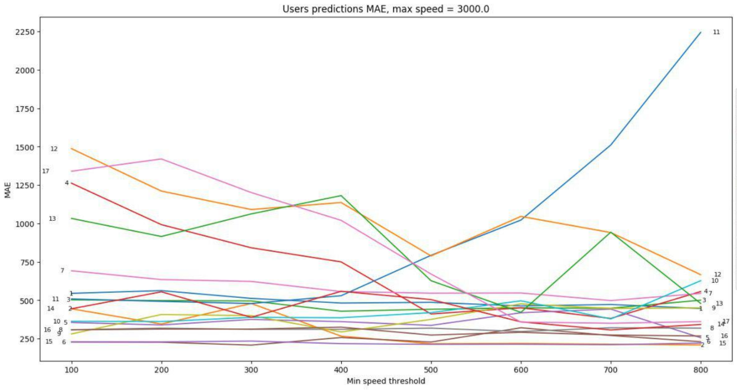

Fig. 4: Dependency of MAE on the lower speed thresholds. UserIDs are at the beginning/end of the curves. All values are for *combined* feature set and *ordered* train/test split.

interpretation of the readers' behavior/actions based on the log data. We also plan to significantly expand the set of textual and contextual features.

Despite modest results we have obtained so far, we believe that the study of reading behavior and users' interactions with text documents enriches representations based solely on content analysis and can be beneficial for various domains. The results of the research may be of interest for school education, reading promotion, book recommendations, and creative writing.

Acknowledgements. We thank *Bookmate* for granting access to the dataset.

References

1. Ralf Biedert, Jörn Hees, Andreas Dengel, and Georg Buscher. A robust realtime reading-skimming classifier. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 123–130, 2012.
2. Anastasia Bonch-Osmolovskaya and Daniil Skorinkin. Text mining War and Peace: Automatic extraction of character traits from literary pieces. *Digital Scholarship in the Humanities*, 32(suppl_1):i17–i24, 2016.
3. Pavel Braslavski, Valery Likhosherstov, Vivien Petras, and Maria Gäde. Large-scale log analysis of digital reading. In *Proceedings of the Association for Information Science and Technology*, volume 53, pages 1–10, 2016.
4. Anne-Marie Brouwer, Maarten Hogervorst, Boris Reuderink, Ysbrand van der Werf, and Jan van Erp. Physiological signals distinguish between reading emotional and non-emotional sections in a novel. *Brain-Computer Interfaces*, 2(2-3):76–89, 2015.
5. Ekaterina Chmykhova, Denis Davydov, and Tatiana Lavrova. Experimental study of factors of speed reading (Eksperimental'noe issledovanie faktorov skorosti chteniya). In Russian. *Psyhologiya Obucheniya*, (9):26–36, 2014.

6. Marios Constantinides, John Dowell, David Johnson, and Sylvain Malacria. Exploring mobile news reading interactions for news app personalisation. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 457–462, 2015.
7. William H DuBay. The principles of readability. `http://www.impact-information.com/impactinfo/readability02.pdf`, 2004.
8. Kai Kunze, Katsutoshi Masai, Masahiko Inami, Ömer Sacakli, Marcus Liwicki, Andreas Dengel, Shoya Ishimaru, and Koichi Kise. Quantifying reading habits: counting how many words you read. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 87–96, 2015.
9. Dmitry Lagun and Mounia Lalmas. Understanding user attention and engagement in online news reading. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 113–122, 2016.
10. János László and Gerald C Cupchik. The role of affective processes in reading time and time experience during literary reception. *Empirical Studies of the Arts*, 13(1):25–37, 1995.
11. Alberto Manguel. *A History of Reading.* Knopf Canada, 1996.
12. Michael E J Masson. Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):400–417, 1982.
13. Victor Nell. The psychology of reading for pleasure: Needs and gratifications. *Reading Research Quarterly*, 23(1):6–50, 1988.
14. Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195, 2008.
15. Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.
16. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
17. Susanne Trauzettel-Klosinski and Klaus Dietz. Standardized assessment of reading performance: the new international reading speed texts IReST. *Investigative ophthalmology & visual science*, 53(9):5452–5461, 2012.