Wordometer Systems for Everyday Life

OLIVIER AUGEREAU, CHARLES LIMA SANCHES, and KOICHI KISE, Osaka Prefecture University
KAI KUNZE, Keio University

We present in this paper a detailed comparison of different algorithms and devices to determine the number of words read in everyday life. We call our system the "Wordometer". We used three kinds of eye tracking systems in our experiment: mobile video-oculography (MVoG); stationary video-oculography (SVoG); and electro-oculography (EoG). By analyzing the movement of the eyes we were able to estimate the number of words that a user read. Recently, inexpensive eye trackers have appeared on the market. Thus, we undertook a large-scale experiment that compared three devices that can be used for daily reading on a screen: the Tobii Eye X SVoG; the JINS MEME EoG; and the Pupil MVoG. We found that the accuracy of the everyday life devices and professional devices was similar when used with the Wordometer. We analyzed the robustness of the systems for special reading behaviors: rereading and skipping.

With the MVoG, SVoG and EoG systems, we obtained estimation errors respectively, 7.2%, 13.0%, and 10.6% in our main experiment. In all our experiments, we obtained 300 recordings by 14 participants, which amounted to 109,097 read words.

CCS Concepts: •**Human-centered computing** → *Interaction devices; Ubiquitous and mobile computing;*

Additional Key Words and Phrases: Eye tracking, reading analysis, wordometer, machine learning, electrooculography

ACM Reference format:
Olivier Augereau, Charles Lima Sanches, Koichi Kise, and Kai Kunze. 2017. Wordometer Systems for Everyday Life. *PACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 123 (December 2017), 21 pages.
DOI: 10.1145/3161601

1 INTRODUCTION

Recording knowledge and experiences has been always essential for humans. With lifelogging and quantified self movements, more and more of this recording can be automated [1, 2]. So far, logging has focused mostly on physical activities such as step counting and food intake [3, 4]. However, there is also a gradual movement toward logging states of the mind, e.g. sleep quality and alertness [5, 6]. In this paper, we deal with a specific mind activity: reading. We aimed to quantify how many words a user reads. Reading is one of the most important cognitive activities in everyday life. Indeed, it has been demonstrated that critical thinking skills and general knowledge are linked to the amount people read every day [7]. That is why in this study, we chose to focus on analyzing reading activity and—more specifically—to quantify it.

This research was conducted in the context of the "reading-life log" [8]. The aim of the reading-life log is to record and analyze everyday reading life. By combining eye tracking with the content of what is read (i.e., what

This work is supported in part by the JST CREST Grant Number JPMJCR16E1, and JSPS KAKENHI Grant Numbers 25240028, 15K12172, and 16K16089.
Author's addresses: Olivier Augereau and Charles Lima Sanches and Koichi Kise, Osaka Prefecture University, Graduate School of Engineering; Kai Kunze, Keio University, Graduate School of Media Design.

2474-9567/2017/12-ART123 $15.00
DOI: 10.1145/3161601

is displayed on a screen) it is possible to record which words are read in conjunction with a time stamp. With this information, many applications are possible: producing a summary about what is read in a day; producing a system to investigate what has previously been read; computing statistics about the kind of documents read; how much is read; how fast a language is read; and which new words are read in a day.

Counting the number of words read has previously been proposed [9, 10]; however, the evaluations were conducted either using professional eye trackers or in a laboratory setting and with small datasets. No study has focused on analyzing reading activity in everyday life, and no device currently available on the market is able to measure such activity. To bring research closer to real-life conditions, we propose in this paper a system for monitoring reading activity and estimating the number of words read using several devices and evaluation methods.

Compared with previous work, the main contributions of this study may be summarized as follows:

(1) we built a large dataset of reading activity representing around 100,000 read words using five eye trackers (section 5.1);
(2) we propose and test a new evaluation model (document and user independent) that is appropriate for application in everyday life;
(3) we demonstrate the usability of inexpensive eye trackers by obtaining an approximate average of 11% error of estimation (section 5.1);
(4) we test the Wordometer systems against complex reading patterns, such as skipping and rereading (section 5.3), and we show that current Wordometer systems are robust to rereading but not skipping.

As noted above, our intention is not to propose another variant of the Wordometer. We aimed to evaluate extensively existing Wordometer systems under various conditions to determine whether the Wordometer may be usefully applied in everyday life.

We used three different devices for our large-scale experiment: one stationary video-oculography (SVoG) system; one mobile video-oculography (MVoG) system; and one electro-oculography (EoG) system. These eye trackers are non-professional systems that may be use in "everyday life situations". They are less expensive but also less accurate than professional eye trackers; thus, the algorithms for the Wordometer have to be designed carefully. Our second experiment focused on analyzing the performance of the Wordometer with complex reading patterns; that is quite important in assessing the application of the Wordometer in real-life conditions. Altogether, we used five devices are used: two SVoGs, two MVoGs and one EoG. In total, 14 users participated in all the experiments, producing 300 recordings which amounted to 109,097 read words.

2 FUNDAMENTAL NOTIONS ABOUT EYE MOVEMENT IN READING AND WORDOMETER SYSTEMS

To allow a proper understanding of related work studies in full detail, we present in this section a few basic concepts about eye gaze analysis.

When a person reads, the eyes have very specific movements. They do not move continuously along a line of text but stop at certain points to process the information from the text (called a fixation) and then jump to another position (called a saccade) [11]. Because of the shape of the parafovea, the horizontal perceptual span is greater than a short word; thus, some words can be guessed and skipped (and have no fixation) while reading. Conversely, some long words can have multiple fixations. After the reader reaches the end of a text line, a large backward saccade going to the next line can be observed; this corresponds to what we call a line break.

The principle of all Wordometer systems is based on the three following steps: (1) processing the signal from the eye tracker; (2) extracting features from the eye movement; and (3) applying machine learning to estimate the number of read words.

Table 1. Three evaluation models used to test the Wordometer. If the learning is user dependent, it cannot be document dependent: the participants never read the same document twice in state-of-the-art (and our) experiments.

Document \ User	Dependent	Independent
Dependent	No	Yes
Independent	Yes	Yes

Machine learning algorithms use a dataset for learning which can be defined in various ways. As shown in Table 1, three different evaluation models can be used to test the performances of the Wordometer:

- User and document independent: to determine the number of read words from the recording of one participant, all recordings of other participants reading different documents are employed.
- User dependent and document independent: the estimation is based on the other recordings of the same participant.
- User independent and document dependent: the estimation is based on all recordings of all other participants (including ones reading the same document and other documents).

Depending on the application and context, different learning methods are appropriate. With a "cold start", no other recordings from the same participant are available; thus, only a user-independent approach can be adopted. If we are considering a group of learners reading the same documents, a document-dependent approach can be applied. Finally, if the reader uses the system for a long time, a user-dependent approach can be employed, and it will better fit the person's behavior.

The performances of the systems are mainly computed in three ways: the average error per recording, A; the weighted average error per recordings W; and the average accumulated error per reader C. The weight used to compute the weighted average is the number of words contained in a document. In our evaluation, the average is always weighted; however, it is not always clear with state-of-the-art whether the error was weighted. A and W are based on the absolute value of error but C is based on the value of error, which can be positive or negative, depending on whether the number of words was over-estimated or under-estimated (details in section 5.1.4). C is used to show the behavior of the prediction for a long recording of one participant; it tends to decrease since the errors compensate one an other over time.

Let E_{pd} be the error of participant p reading document d, P the number of participants, D_p the number of documents read by participant p, and N_d the number of words of document d. The three measures A, W, and C are defined as follows:

$$A = \frac{\sum_{p=1}^{P} \frac{\sum_{d=1}^{D_p} |E_{pd}|}{D_p}}{P},$$

$$W = \frac{\sum_{p=1}^{P} \frac{\sum_{d=1}^{D_p} |E_{pd}| \times N_d}{\sum_{d=1}^{D_p} N_d}}{P},$$

$$C = \frac{\sum_{p=1}^{P} \frac{\sum_{d=1}^{D_p} E_{pd} \times N_d}{\sum_{d=1}^{D_p} N_d}}{P}.$$

It is evident that in our definition, C is also weighted by the number of words in the documents. However, it is unclear if that is the case in the state-of-the-art systems.

3 RELATED WORK

First, we present research related to the general topic of eye movement and reading analysis. Then, we present previous studies about Wordometer systems and their limits.

3.1 Eye movement and reading analysis

Three main technologies are commonly used to analyze the movement of the eyes: eye-attached tracking, such as scleral search coils (SSC), electro-oculography (EoG), and video-oculography (VoG).

SSC is regarded as a gold standard for recording accurate data of eye movements in terms of resolution, accuracy, low noise and fast response [12]. But it is the least usable in everyday life condition: the subject needs a topical anesthetic for insertion of the SSC in the eye, and a magnetic frame is placed around the subject.

EoG has been especially used for human-computer interaction [13] and activity recognition [14]. The data recorded with the EoG are not very accurate and tend to be noisier than with other approaches: facial muscles and head movements interfere with the eye movement signal [15]. Thus, EoG systems have not been commonly used for reading analysis except for a few exceptions such as Kunze et al. [16] and Ishimaru et al. [10].

VoG is one of the most popular eye tracking technologies. There are two types of VoG: mobile ones (MVoG), which are head mounted, and stationary ones (SVoG), which are usually attached beneath a computer screen. Several applications have been developed for such applications as activity recognition [17, 18], detecting user interest [19], and biometric identification [20].

Further, VoG is the main system used for reading analysis and extracting information about readers and documents. From the reader's perspective, some applications concern determining the reader's comprehension [21], English ability [22], and the TOEIC (Test of English for International Communication) score [23]. From the text perspective, the eye gaze can be used to assess the quality of a text [24] and categorize documents [25] in addition to other applications. Some studies have focused on specific types of documents, such as sheet music [26, 27] and comics [28].

Eye movement while reading was analyzed by Rayner around 40 years ago [29]. Informative features about fixations and saccades are extracted from the eye tracker signal to analyze the reading behavior.

3.2 Existing Wordometer systems

In this section we present the three main existing Wordometer systems.

The first publication about the Wordometern by Kunze et al., appeared in 2013 [9]. The authors used the SMI mobile eye tracker[1] (MVoG). The world camera of the mobile eye tracker is used for document image retrieval and correcting the 3D perspective transformation. In the experiment by Kunze et al., nine subjects read 10 documents each. The 90 recordings represented a total of 27,930 words read. For each recording, five features were extracted: the duration required for reading, number of fixations, total distance of eye movements, total distance of saccades, and average distance of saccades. The evaluation model tested by the authors was user independent but not document independent. In order to estimate the number of read words of the recordings of one participant, all the recordings of all other participants were used to build a support vector regression model. The drawback of not using a document-independent model is that to determine the number of read words in a document, some other people must have read the same document; that is not so probable in real-life conditions. Then, the errors in each document are totaled, which compensates for the under-estimation and over-estimation of some recordings. Finally, according to the Kunze et al., they obtained an average error of 8.2% for each participant. The details of their were provided, and we computed the weighted average error per recording; we obtained an error of 14.6%.

In 2015, Kunze et al. [16] presented an updated version of the Wordometer. That time, the number of read words was estimated directly from the eye gaze data recorded by the mobile eye tracker (without using the recorded

[1] http://www.eyetracking-glasses.com/

image of the document from the scene camera); the authors used a medical EoG and a prototype version of a commercial EoG called JINS MEME[2]. Kunze et al. conducted four experiments in that study. The first involved reading from paper with a mobile eye tracker; 14 documents were read by nine subjects. Three of the five features from the previous study are different. In this study, the features employed were as follow: total time reading, sum of all saccade distance, sum of the line break saccade distances, number of line breaks, and sum of the reading saccade distances. The authors used a similar approach as in their previous study (user independent, document dependent, support vector regression), and obtained an 8% error for each participant. The authors did not provide complete details, and so we cannot compute the weighted average error per recording. The second experiment involved reading from several screens of different sizes (e.g., e-ink reader, smartphone). Five documents were read by 10 subjects. The average error using the same approach was 17%. Kunze et al. considered that the error was largefi?!especially when the participants had very long reading lines or when they moved their heads excessively. The third experiment was the same as the second except that a medical EoG was used. In that experiment, eight participants read five documents. The authors obtained an average error of 5%. The fourth experiment was undertaken with JINS MEME and four participants. Unfortunately, the number of documents and read word was not detailed in the paper. The authors obtained an average error of 20%.

More recently, Ishimaru et al. introduced the Wordometer 2.0 based on JINS MEME [10]. Five participants read 38 paragraphs which amounted to 190 read paragraphs or approximately 10,000 read words. Four features were extracted from the EoG signal: total number of forward saccades, mean EoG signal value of forward saccades, total number of backward saccades, and mean EoG signal value of backward saccades. Those features were then inputted into a support vector regression algorithm to predict the number of read words. An average error of 18% per paragraph was obtained in a user-independent (but not document-independent) approach.

3.3 Limitations of the state-of-the-art systems

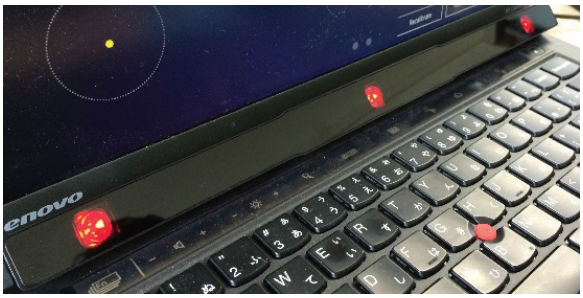
Hitherto, only three devices have been used to apply the Wordometer: the SMI mobile eye tracker (MVoG); a medical EoG; and a prototype version of the JINS MEME (EoG). That prototype was used in previous studies [10, 16] works at 11Hz, and is based on different hardware from what we used. That device is not commercially available; thus, we decided to use the Academic version; that is commercially available[3], but it works differently (sampling rate, 100Hz).

In our experiments, we used five eye tracking systems including the three following systems that have not yet been used: Tobii Eye X[4] (SVoG), SMI RED250[5] (SVoG), and Pupil[6] (MVoG), and a new version of JINS MEME (EoG).

Only two state-of-the-art models have thus far been evaluated: user-dependent and document-independent learning; and user-independent and document-dependent learning. In the present study, we investigated another evaluation model, which is relevant to real-life usage: user-independent and document-independent learning. This method corresponds to the situation where the document read by other readers and used as a learning dataset is different from the document used as a test; that situation is likely to occur in real-life conditions.

To summarize, the main differences with previous studies are as follows: the focus on using inexpensive eye trackers; the creation of a large dataset; the proposition of a new evaluation model; and the test against complex reading patterns.

[2]https://jins-meme.com/en/
[3]https://jins-meme.com/en/academic/
[4]https://tobiigaming.com/
[5]http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/red250-red-500.html
[6]https://pupil-labs.com/pupil/



(a) Tobii Eye X stationary eye tracker, one of the most inexpensive eye trackers on the market.

(b) SMI Red 250 stationary eye tracker with high frequency and accuracy.

Fig. 1. The two stationary eye trackers (SVoG) used for our experiments.

We would like to point out that the proposed Wordometer systems are based on the previous work [10, 16]. We directly used those algorithms but made minor changes to optimize the performance with the current data and devices. We will not compare our results with the previous algorithms since they give slightly worse results.

4 PROPOSED WORDOMETER SYSTEMS

In this section we present the five devices used in our experiments.and the corresponding algorithms to determine the number of read words. An essential part of the Wordometer algorithm involves identifying and quantifying the fixation and saccade features. The general framework is similar to the Wordometer related studies [9, 10, 16]; however, we made minor changes such as changing the filtering and adapting some features to optimize the performances for everyday life eye trackers.

We provide explanations of the systems, including the minor changes, in the following sections.

4.1 Stationary video-oculography

The stationary eye tracker is attached beneath a computer screen, as illustrated in Fig. 1. The raw eye gaze positions and time stamp are recorded by the system while the user is reading on the screen.

With our system, the Wordometer algorithm comprises the following four main steps:

(1) detecting fixations and saccades;
(2) detecting line breaks;
(3) extracting features;
(4) applying a regression.

Detecting fixations and saccades. The raw signal is processed using the algorithm of Buscher et al. [30]: when the eye gazes are concentrated on a small area for a sufficient time, a fixation is detected. The rapid eye movement between two fixations is represented as a saccade. Figure 2 shows an example of fixations and saccades based on Tobii Eye X raw data.

Detecting line breaks. The sequence of fixations and saccades is processed to determine line breaks, which correspond to a large eye gaze regression: when the reader goes back from the end of one text line to the beginning of the next. The number of line breaks can differ from the number of text lines if the reader rereads or skips some parts of the text.

To identify the line breaks, we first merged the consecutive backward saccades. This step helped differentiate between small backward saccades (corresponding to rereading a word or noisy saccades) and long backward saccades (corresponding to a line break).

We the computed the average length of all the backward saccades. We experimentally fixed a threshold α; a backward saccade was detected as a line break if its length B validated the corresponding condition:

$$B > \alpha \times \sum_{i=1}^{N} B_i,$$

where N is the number of saccades.

Extracting features. We extracted five features are extracted for each recording:

- total reading time;
- sum of line break saccade distances;
- sum of all saccade distances (without the line breaks);
- number of line breaks;
- number of fixations.

We computed a saccade distance as the distance between two fixations. The large regressive saccades detected as line breaks are counted separately. These features are partially based on those introduced in [9, 16].

Applying a regression model. We attempted to predict the number of read words as the number of words contained in the text we asked a participant to read. Thus, for each reading session, we used the number of words the text contained as the ground truth even if there were slight differences with that number (for example if the reader skipped or involuntarily reread a few words). However, participants were asked to follow one of three specific scenarios: (1) no rereading or skipping any part of the text; (2) reread a given paragraph; and (3) skip a given paragraph. Accordingly, in each case, we knew the number of words the participants were supposed to have read, which we used as the ground truth.

We used the five previously extracted features as the input for a SVR (Support Vector Regression) model. We employed three different evaluation models, as detailed in section 2. We determined the number of read words in one recording based on the recordings used as learning data.

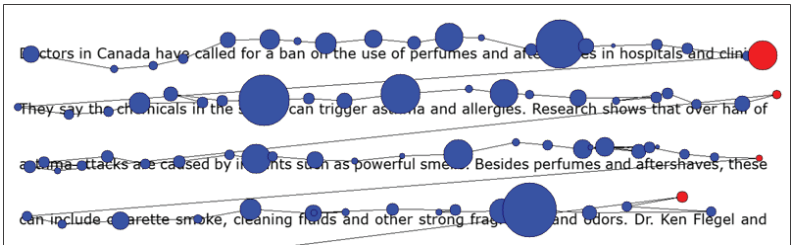
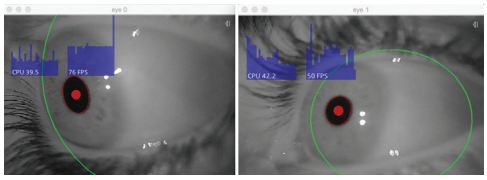


Fig. 2. Example of data obtained after processing the Tobii Eye X raw data with the Buscher et al. [30] algorithm. The blue circles represent fixations. The diameter of the circles indicates the duration of the fixations. The lines between the circles represent saccades. The last fixations before a line break are colored red.

(a) SMI mobile eye tracker, a professional eye tracker.

(b) Pupil mobile eye tracker, an open source and inexpensive device.



(c) Image recorded and processed by the eye cameras of the Pupil eye tracker on a European participant.

Fig. 3. The two MVoG eye trackers used in the experiments: SMI mobile and Pupil.

4.2 Mobile video-oculography

The mobile eye tracker is a headset comprising three cameras: two record each eye; one front camera (also called the scene or world camera) records the scene. The two MVoGs that we used appear in Fig. 3: the SMI and Pupil mobile eye trackers.

The algorithm of the Wordometer is very similar to that used for SVoG. But the Buscher et al. fixation and saccade algorithm does not perform optimally with a mobile eye tracker. The fixed-size area used for detecting fixations is defined in the screen coordinates; but the mobile eye tracker records fixations in the scene image coordinates, as shown in Fig. 4. The screen displaying the text is included in that image; thus, its size changes depending on whether the participant is closer or farther from the screen. Accordingly, we used another algorithm based on gaze dispersion to detect the fixations and saccades [31]. The code is open source and available on GitHub[7].

After extracting the fixations and saccades, the next parts were exactly the same as for the SVoG. First, we tried to detect the line breaks. To do so, we merged the consecutive backward saccades. Then, if a backward saccade was longer than a fixed threshold α, a line break was detected. We used the same five features as for the SVoG: total time reading; sum of line break saccade distances; sum of all saccade distances, number of line breaks, and number of fixations. We then applied a regression model to determine the number of read words based on these features.

[7]https://github.com/pupil-labs/pupil/blob/master/pupil_src/shared_modules/fixation_detector.py

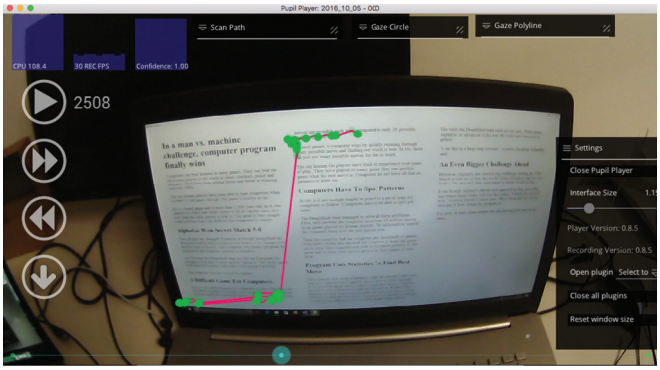
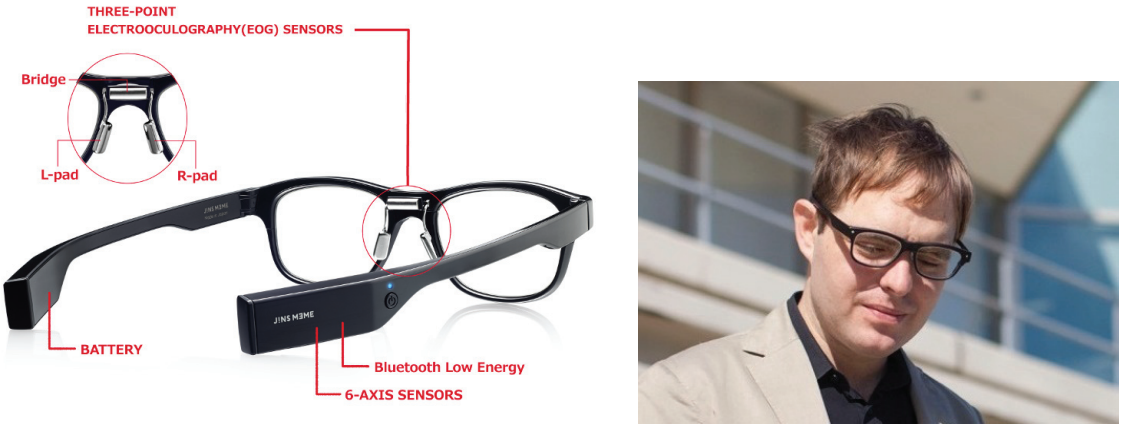


Fig. 4. Image recorded by the world camera, representing what the user sees. The fixations appear in green and the saccades in pink. The image is used for illustration but is not used by the algorithm.



(a) JINS MEME glasses.

(b) JINS MEME glasses as worn by a user.

Fig. 5. JINS MEME glasses with EoG sensors.

4.3 Electro-oculography

The EoG system we used comprised three electrodes included in the frame of the glasses. The glasses are depicted in Fig. 5. We used the Academic version of JINS MEME, which is commercially available.

EoG involves measuring the potential between the front and back of the eye. When the eye is moving, the electrodes measure the potential which is proportional to the angle.

The main steps of the Wordometer are the same as for SVoG. However, the signal processing and features are different since the signal is of a different nature (Fig. 6).

First, we applied a Butterworth filter to smooth the signal. Then, we processed the signal with a 3-second window. If the standard deviation of the windows was four times greater than the standard deviation of the whole signal, the window was regarded as too noisy and ignored in the next part of the algorithm. Some recordings

[8]https://github.com/jins-meme/ap-datalogger-for-windows

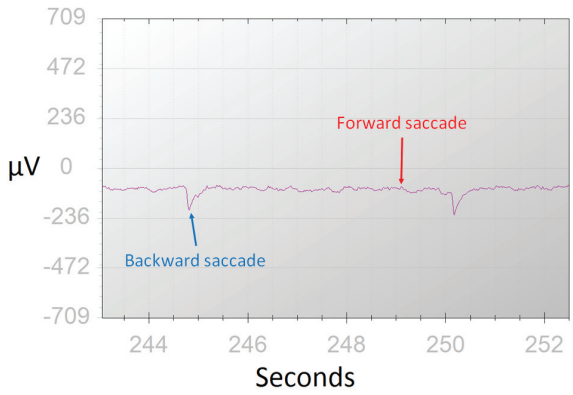


Fig. 6. Example of an electro-oculogram extracted using the JINS MEME Data logger software[8]. Two large backward saccades representing line breaks are evident. Several small forward saccades are also visible.

contained too much noise: they were entirely removed from the experiment dataset and not included in the results.

We extracted the same four features as the authors of [10]. All the local maximum values were detected and considered to be forward saccades. We defined two features made from the number of local maxima and average value of the maxima. The large backward saccades representing the line breaks were detected as the local minimum values; they were lower than an experimentally defined threshold. For better performance, we set the threshold in a different manner than in [10]. In a 1-second window, a local minimum is detected as a large backward saccade if:

$$L < A - 3 \times S,$$

where L is the value of a local minimum, A is the average value of the window, and S is the standard deviation of the window. This is done to accept only the large regressions that correspond to line breaks. Then, two other features are extracted: the number of minima and average value of the minima. Among these four features, we found that the two first ones were not very relevant. The signal of the EoG was slightly noisy; thus, it was very difficult to distinguish the small forward saccade from the noise. These results were improved by using only the last two features which are more robust (based on the large backward saccades).

5 EXPERIMENTS

In the first experiment, we estimated the number of read words by the participants reading a text displayed on a screen, using the everyday life eye tracking systems: Tobii Eye X, Pupil, and JINS MEME Academic version (which gives access to raw signals with a sampling frequency of 100 Hz). We experimentally fixed the threshold to detect the line break at: $\alpha = 0.75$. After presenting the main experiment results in section 5.1, we will detail the main causes of error for the Wordometer system in section 5.2.

In section 5.3, we analyze the robustness of the Wordometer with respect to special reading behaviors, such as rereading and skipping: that is important in approximating real-life conditions.

The documents used in all experiments were extracted from the website Newsela [9]. That website provides news with different grades of English difficulty. We selected texts with different English levels that would be

[9]https://newsela.com/

appropriate for our the readers: not too easy and not too difficult for undergraduate or graduate non-native English speakers.

In all experiments and with all devices, the participants were free to move their heads and bodies and to make any natural motions. We just asked them to sit in front of a computer screen to read a text. The participants were requested to press a key on a keyboard to indicate when they were starting to read and press it again when they were stopping. Those time stamps were used to segment the recorded signals.

5.1 Reading with everyday life eye trackers

The aim of this large-scale experiment was to demonstrate whether inexpensive eye tracking devices could be used to determine the number of words read in daily life. The eye trackers were the Tobii Eye X (SVoG) Pupil (MVoG) and JINS MEME (EoG). Since the SVoG can be used only with a screen, we decided to use that procedure (i.e. to display documents on a screen) with the other devices to achieve a fair comparison.

To ascertain exactly the number of words read by the subjects, we asked them to read the text from beginning to end without rereading or skimming any parts of the text. Despite this instruction, some participants could not refrain from rereading some words for the sake of understanding or skipping some words if they found the text too easy. Still, we considered that the number of words contained in the text was the number read by the subjects. A natural reading pattern also involves rereading and skimming; thus, for some specific texts, we asked some participants to reread or skip one specific paragraph. That process will be explained for the second experiment.

In this experiment, 14 people participated. All the subjects were university students (11 males, three females). The English ability of the participants varied and they were of different nationalities (Japanese, French, Malaysian, Norwegian, and Chinese). Most of the subjects were undergraduate or graduate university students; they were aged 21fi??25 years. Recording while wearing glasses is not possible with the JINS MEME and Pupil; thus, we selected mainly participants who used contact lenses or did not wear glasses. Only one subject wore glasses, and he took part in the experiments only with the Tobii Eye X.

In this experiment, we made 194 recordings. Not all the participants were willing to spend a long time with the experiment, so they read different numbers of texts. In total, the subjects read 78,579 words. With all the participants, all the texts were read for the first time and read only once.

Table 2 shows the results of the experiment. The performances were analyzed based on three different learning processes, which are detailed in the following sub-sections. We refer to the "weighted average" as the average weighted by the number of words contained in the text of a recording. We also computed the weighted standard deviation and the weighted median, defined as follow. Let μ be the weighted average, w_i the weight of an element x_i, and N is the number of elements. The weighted standard deviation S_w is:

$$S_w = \sqrt{\frac{\sum_{i=1}^{N} w_i (x_i - \mu)^2}{\sum_{i=1}^{N} w_i}}$$

The weighted median is found as the element x_k satisfying both of these conditions:

$$\begin{cases} \sum_{i=1}^{k-1} w_i < 1/2 \\ \\ \sum_{i=k+1}^{n} w_i \leq 1/2 \end{cases}$$

For each pair of evaluation models, we attempted to determine if the difference in performance was statistically significant using a paired t test. The P values were less than 0.05 for the three pairs of models for the EoG but not for the other devices. The evaluation models used different learning data, so the results may be significantly different in some cases. Specifically, the learning data for user-dependent learning is much smaller than for user-independent cases, which tends to produce worse performance.

Table 2. Weighted average, standard deviation, and median error in percentages for the Wordometer with each device based on the different learning approaches: user and document independent, user dependent, and user independent. *There were 14 participants and some participants used several devices. **There were 18 documents and the same documents were used for different experiments. The SVoG and the EoG were recorded together.

	Device	SVoG	MVoG	EoG	Total	
	Participants	9	10	5	14*	Weighted Av.
	Documents	18	18	18	18**	error
	Recordings	82	53	59	194	among devices
	Words	33249	20734	24596	78579	
User & doc. independent	Weighted Av. Error	14.0%	7.6%	11.3%		11.5%
	Weighted Std Error	10.2%	7.3%	8.4%		
	Weighted Median Error	11.3%	4.6%	9.0%		
User dependent & doc. indep.	Weighted Av. Error	11.8%	5.6%	9.5%		9.4%
	Weighted Std Error	9.1%	5.5%	7.9%		
	Weighted Median Error	9.6%	4.3%	7.9%		
User independent & doc. dep.	Weighted Av. Error	13.0%	7.2%	10.6%		10.7%
	Weighted Std Error	9.5%	6.8%	8.1%		
	Weighted Median Error	10.9%	5.1%	8.7%		

The participants and documents were partially different, so to compare the performance of the devices we used the Welch t test. We tested all three pairs of devices for the user- and document-independent models. We found a significant difference (P values less than 0.05) between the SVoG and EoG. For the other cases, the P values were greater than 0.05, which means that no conclusions could be drawn.

In the following sub-sections, we detail the results based on the three evaluation models in terms of participant and documents. After that, we conduct an analysis of the sources of errors.

5.1.1 User- and document-independent learning. We wished to address a problem that has not hitherto been considered in studies about the Wordometer: user and document independence, which is the most difficult context. No document-independent learning has been tested with state-of-the-art systems. To determine the number of read words in the recording by one participant, we used all the recordings of all other participants reading a different document.

In our database, several recordings corresponded to different subjects reading the same documents. If the reading behavior of two participants was similar, the number of read words by one subject would be easy to estimate: simply using the recording of another participant reading the same document. But in a real-life application, if different people used the Wordometer for reading different texts, there would be a low chance for two people to read the same document.

The weighted average errors with the MVoG, SVoG, and EoG were all under 15%, which led us to conclude that such systems could be used in everyday life. As a comparison, the error with pedometers used in wearable devices was found to vary from 1.5% to 22.7% [32].]. If we employed a similar daily-life approach for one subject, the errors of that personfis recordings would tend to compensate one another: some errors would be

Table 3. Cumulative error of the estimation per participant in user- and document-independent learning. Each subject undertook as many recordings as they wanted with one, two, or three devices.

Participant	SVoG		MVoG		EoG	
	# of words	cumulative error	# of words	cumulative error	# of words	cumulative error
R1	2201	15.9%	4454	9.7%		
R2			238	48.4%		
R3	5712	9.1%	690	1.1%	4378	2.1%
R4			5733	5.5%		
R5			1440	9.8%		
R6			803	0.9%		
R7	1346	7.4%	2690	3.4%	4608	9.2%
R8	5097	12.9%	1729	8.2%	5097	9.6%
R9			872	0.5%	6426	9.3%
R10	5213	1.0%	2085	1.7%		
R11	4656	6.6%			4087	8.5%
R12	4219	3.8%				
R13	2642	12.4%				
R14	2163	23.6%				
Sum / W. av.	33249	9.0%	20734	6.2%	24596	8.0%

positive (overestimation); others would be negative (underestimation). We will show in the participant-centered experiment result (section 5.1.4) that the cumulative error was two to three time lower than the average error.

5.1.2 User-dependent and document-independent learning. We computed the performances of the system in a user-dependent way. In this case, only the other recordings of the same subject were used to estimate the number of read words for one recording. This method can be applied only if the participant is trained in the system before using it. The details of the results appear in the second row of Table 2.

It is evident that the performances obtained with this learning approach were the best among all three approaches. Indeed, if this model can fit reader behavior, this system can estimate more accurately the number of read words. Thus, the Wordometer can be used either in a user-independent or dependent manner.

5.1.3 User-independent and document-dependent learning. To make a comparison with previous studies, we made an analysis in a user-independent and document-dependent way. The number of words of one recording was determined based on all recordings from all other subjects (including ones reading the same document and other documents). The last row of Table 2 shows the results.

It is evident that in this case the performances are intermediate between user and document-independent approaches and user-dependent approaches. The Wordometer systems performed better if similar documents had been read by other participants. Unfortunately, this situation might not often occur in daily life.

5.1.4 Performances per participant. Errors are sometimes positive (overestimation) and sometimes negative (underestimation); thus, we determined the cumulative error for each participant. This demanded a long-time

Table 4. Five documents with the largest weighted average error among the devices compared with all documents.

Doc ID	Difficulty (L)	Nb of words	Av. error
ALP	900	643	27.7%
SIT	840	253	16.7%
ITA	950	238	16.6%
FIF	860	293	15.8%
FIG	910	485	15.0%
Average 5	892.0	382.4	18.4%
Average All	848.9	405.4	11.3%

reading behavior analysis. We did not conduct this estimation for each text, but after a subject had finished reading all texts. Indeed, the aim of the Wordometer is not to quantify specifically how many words have been read after a short period, such as reading a few words; it is for reading several documents or reading within a day.

We present the results for each subject in Table 3. This table shows the performances of the algorithm for each user according to the eye tracker used in the experiment. It is evident that the cumulative errors are notably smaller than the weighted average error reported in Table 2 for all devices. We observed the best average performances for the MVoG. However, it is clear that depending on the subject, the SVoG can perform better than the MVoG (as with participant R10) or than the EoG (as with participant R7).

It is also evident that some subjects (especially participants R2 and R14) had a very large error. This occurred when the calibration failed or detection of the eyes was not possible. We will explain the sources of error in greater detail in section 5.2.

Since each Wordometer system is independent of other such systems based on technology (SVoG, MVOG, and EoG), we did not request that the participants necessarily used all the devices. Some subjects had limited time to participate in the experiment. Therefore, we decided to record more participants with fewer devices than fewer participants with more devices. In the future, we intend to enlarge our dataset.

5.1.5 Performances per document. We analyzed in greater detail the performances of the systems according to the type of document. The dataset comprised 18 documents: each was read by at least six participants, at most by 14 participants, and on average by 9.4 participants.

Each text was associated with a difficulty based on the estimated Lexile [10] measure [33]. As explained on the Lexile Web site: "a Lexile text measure is based on the semantic and syntactic elements of a text". "For example, the first Harry Potter book measures 880L". In our dataset, the document with the lowest difficulty was 610L, and the text with the highest was 950L. The average difficulty of all the documents was 848.9L. Our participants were mainly university students from the computer science department and were non-native English speakers; thus, we chose texts that were not too complex.

Table 4 shows the five documents with the largest error compared with the other documents. The averages were determined among the three devices. The number of words of each document is also displayed in Table 4. In our dataset, the shortest document contained 238 words and the longest 643 words. On average, the documents contained 405.4 words. It is evident in Table 4 that the documents with greater difficulty produced a higher estimation error than documents that were easier to understand. We found a significant correlation (P value

[10]https://lexile.com/about-lexile/lexile-overview/

less than 0.05) between the performances of the SVoG and the difficulty of the documents based on the Pearson correlation test. Indeed, we obtained the following P values for the three evaluation models:

- User dependent and document independent: $P = 0.018$,
- User independent and document dependent: $P = 0.033$,
- User and document independent: $P = 0.013$.

This result means that if the document was too difficult, the reading behavior of the participants changed and the estimation became less accurate.

It is also evident that on average, if the documents are shorter, the number of read words is more difficult to predict. This is because fewer features are available to make the prediction; thus, a slight noise in the signal will have a greater impact on the prediction.

However, one exception occurred for the document with the largest number of words; that was also the document with the greatest error. As participants needed to move their heads more while reading, that would result in greater noise in the eye-tracker signals.

5.1.6 Comparison with professional eye trackers. Professional systems have a higher sampling rate and accuracy, but they are more expensive [34]. However, we wanted to demonstrate that for simple applications, such as counting the number of read words, very high sampling rate and accuracy is not necessary. Furthermore, our algorithm was designed to be robust to noise and some accuracy errors. The preprocessing step (computing the fixations) and detection of line breaks are designed to be simple and easily appliedfi?!even with an inexpensive device.

We used two professional eye trackers to test the Wordometer: the SMI Mobile (MVoG) and SMI Red 250 (SVoG). The dataset and results of the experiments appear in Table 5, which shows the performances of all devices used for all the experiments. It is clear that the performances of the professional devices were not higher than the everyday-life devices for counting the number of words. That led us to conclude that our Wordometer systems operated well as everyday-life eye trackers.

Table 5. Percentage of weighted average error and cumulative error per participant with the Wordometer under user- and document-independent learning.

Device type	Device name	Approx. Price (USD)	Participants	Rec.	Words	Weighted av. error	Cumulated error
MVoG	SMI Mobile	11,900	4	14	4556	13.0%	6.7%
MVoG	Pupil	1,500	10	53	20734	7.6%	6.2%
SVoG	SMI Red250	23,000	4	14	4881	15.8%	10.8%
SVoG	Tobii Eye X	110	9	82	33249	14.0%	9.0%
EoG	JINS MEME	2,250	5	59	24596	11.3%	8.0%

5.2 Analysis of errors

Before presenting the next experiment, we analyze here the main sources of errors.

5.2.1 Calibration. Calibration is an essential part of the protocol before starting to use an eye tracker. However, depending on the device, it can be relatively easy or difficult to check whether or not the calibration was successful.

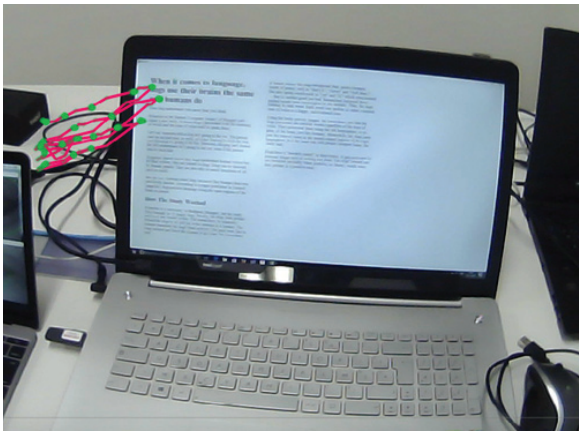


Fig. 7. Example of a recording where the calibration failed for participant R8 while using the MVoG. This recording resulted in an estimation error of 27.3%.

In this experiment, the most difficult eye tracker to calibrate was the MVoG. As evident in Fig. 7, if the calibration is not conducted properly, the recorded position of the eye gaze will be very different from the real position.

5.2.2 Head movement. With most eye-tracking systems, head movements result in a less accurate estimation of the number of read words. Head and facial muscle movements produce noise in the EoG signal, which is difficult to differentiate from eye movement since the amplitude is the same as the signal itself. Indeed, if the eyes are fixed on a position in a document while the head moves, the EoG will detect that the eyes are moving (which is true relative to the head, but not relative to the document).

The SVoG is attached to the screen, so if the subject moves their head too much, the eye-tracking position will be inaccurate. In our previous results, we found that the MVoG performed better than the other systems; one reason for this finding is that the MVoG is more robust regarding head movements of the user.

5.2.3 Pupil detection. Pupil detection is the main problem for the SVoG and MVoG systems. It can affect calibration and recording during an experiment. With the MVoG, the eye cameras are placed very near the eyes, and it is not always easy to adjust them for some participants.

With the SVoG, after the calibration is finished, if the participants change their posture too much, pupil detection can be lost. Unfortunately, we had no way of detecting when that happened: the SDK provides no feedback about the accuracy of pupil detection.

During the experiments, we experienced some trouble with pupil detection, especially with some Asian subjects. Indeed, if the eyelid is less widely open, part of the pupil is concealed by the eyelid and cannot be perfectly detected.

Fig. 8 shows the impact of misdetection of the pupil. In some extreme cases, such as for participant *R*2, misdetection of the pupil produced considerable noise and disrupted the proper functioning of the algorithm.

5.2.4 Electrodes. The JINS MEMEfis EoG consists of three electrodes, which have to be in contact with the skin (Fig. 5a). If one of them is not properly touching the skin, there will be a high impact on the recording.

Thus, if the user did not wear the glasses properly or moved them too much during the recording, the recordings tended to be noisier. We did not include five recordings in the dataset because they contained too much noise. We believe that occurred because the electrodes of the glasses were not properly in contact with the skin. To

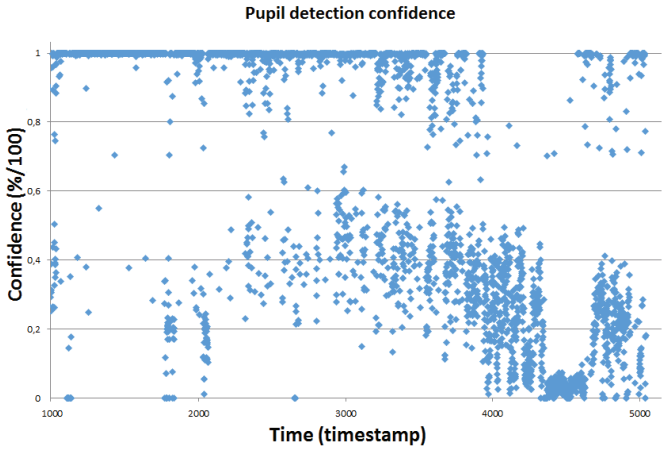


Fig. 8. Example of a recording where the pupil detection confidence was low for participant R2 when using the MVoG. This recording resulted in an estimation error of 48.4%. It is especially evident at the end of the recording (when the reader was looking at a corner of the screen) that the pupil was no longer detected. The more fixations are found at an incorrect position, the greater will be the impact on the performance of the system.

solve that problem, an algorithm needs to be developed that warns the subject to put the glasses correctly in place. However, the advantage of the JINS MEME is that calibration is unnecessary.

Some recordings were partially noisy and were retained in the dataset; the corresponding parts were removed. This is explained in section 4.3.

5.2.5 Nationality, sex, reading skill. We also investigated the impact of nationality (Asian / European) and sex (male / female) on the performance of the Wordometer; however, we found no statistically significant differences.

We also wanted to analyze the impact of the readerfis skill since it directly affects reading behavior. Unfortunately, that necessitated each participant passing a standardized test, which we did not do. We believe the Wordometer would be more accurate if the learning dataset contained only reading data from participants with similar reading skills to the subject being tested. But this point remains to be proven.

5.3 Rereading and skipping behavior

In this experiment, we analyzed the robustness of our systems with different types of reading behavior: rereading and skipping. This experiment is important in quantifying the discrepancy between laboratory conditions and everyday-life conditions; hitherto, it has not been undertaken in research.

Since the baseline is very difficult to obtain in totally free reading, we asked the participants to follow a particular scenario: skipping or rereading a specific paragraph and reading all other parts of the text without rereading and skipping. In this way, we were able to establish the baseline since we could determine the number of read words.

We used the previous dataset for this experiment, and we added some recordings containing rereading and skipping. The reading conditions were the same as in the previous subsection: similar kinds of documents were used, and 10 of the 14 subjects from the previous experiment participated in this one. The details of the new data and corresponding results appear in Table 6.

In this table it is evident that the current version of the Wordometer is not robust to skipping behavior: the saccades can be confused with normal reading behaviorfi?!especially if that occurs only for a short time. That

Table 6. Percentage of error of the Wordometer while rereading and skipping based on user- and document-independent learning. The most problematic is the skipping pattern, especially with the SVoG and MVoG systems.

	Device	Participants	Rec.	Words	Weighted av. er.	Weighted Std	Weighted Med.
Rereading	SVoG	6	24	7876	19.2%	13.6%	17.1%
	MVoG	4	12	4108	6.55%	4.88%	4.97%
	EoG	4	9	2863	19.7%	6.9%	16.7%
	Total	9	45	14847	15.8%	9.90%	13.7%
Skipping	SVoG	5	12	2358	53.8%	37.0%	37.6%
	MVoG	4	13	2470	60.7%	94.9%	25.2%
	EoG	4	8	1406	21.6%	19.2%	14.8%
	Total	8	33	6234	49.2 %	55.9%	27.5%
Total	All	10	78	21081	25.7%	23.5%	17.8%

was especially true for the SVoG and MVoG systems. The EoG was more robust to skipping behavior since only large backward saccades are used in the algorithm.

Furthermore, since the number of read words is smaller, the estimation is more difficult (as explained for short texts in section 5.1.5). Another algorithm for classifying reading and skipping, such as that developed by Biedert et al. [35] needs to be used with the Wordometer.

Rereading behavior is less problematic. Indeed, reading two lines once or reading one line twice produces similar fixation and saccade features and a similar estimation about the number of read words. To be usable in everyday life, detecting rereading behavior is not necessary; however, more examples of these specific patterns should be used to make the algorithm more robust.

6 DISCUSSION AND LIMITATIONS

The Wordometer works well with inexpensive eye trackers in everyday-life usage, but there is still a discrepancy between laboratory conditions and daily life. Among the three devices, the most user-friendly are the EoG glasses. No calibration is needed, and the glasses look like normal ones used in daily life. The SVoG is also quite user-friendly. It is attached beneath the screen and is not intrusive. However, a quick calibration is needed. The MVoG is the most intrusive device: it is not so easy to wear or calibrate.

With all three devices, no feedback is given to the subject during the recording. If the subject moves the glasses or if detection fails, the subject is unaware, and the result of the prediction will be inaccurate. This problem is not specific to Wordometer systems: it applies to any eye tracker. These devices are normally used in laboratory conditions, and so this problem has not yet been addressed.

Our experiments were conducted using computer screens after reading at least 238 words (the size of the shortest text in our dataset). If the screen is too small or reading time is too short, the current systems will not work properly. If there are no line breaks or very few lines, the estimation will become inaccurate. However, the Wordometer is rather like a pedometer: the latter cannot be used for estimations if a person walks just one or two steps. In the same way, the Wordometer will work well if the subject is reading continuously for at least a short period of time.

We demonstrated that the difficulty of the document has an impact on reading behavior and on the performance of the Wordometer. Thus, the difficulty of the document could be used as a parameter to improve performance.

We also believe that the readerfis skill has an impact; that could also be estimated after reading several texts with a method such as that of [23] and used as an input feature. The final and main problem is that another algorithm is necessary to detect the activity of the subject: is that person reading or doing something else? The Wordometer will work only if the user is reading. In our experiments, the recordings were started and stopped manually, but that has to be done automatically. Preliminary work about activity detection with an eye tracker has been conducted [14, 17, 32]; that approach should be extended and integrated with our system. For example, browsing a Web site and skimming text might be detected as reading, but they would be incorrectly evaluated by our system and would need to be filtered. Thus far, no such systems have become available; more research should be undertaken in this direction.

However, we consider that the reported result of an approximate 11% of errors is encouraging and could be sufficient for everyday usage. By way of reference, the accuracy of pedometers is reportedly 12% [36]. Counting the number of words read and determining the number of steps walked are quite different activities. But we have demonstrated a means of quantifying daily mental activity in the same way as it is possible to quantify physical activity.

7 CONCLUSION AND FUTURE WORK

Quantified self-movement is becoming popular: to better understand themselves and change their behavior, people track their daily activities, such as doing sports or eating. The Wordometer we present in this paper is a system that counts the number of words read without analyzing the content of what is read. We believe this approach will help people understand more about their daily cognitive activity and encourage them to read more.

To determine the number of words read by a subject, we presented in this paper two algorithms for processing VoG and EoG signals. With the VoG system, if the calibration failed or the participantfis eyes were not correctly detected, we obtained a large error. The EoG system is more sensitive: the three electrodes always have to be in contact with the subjectfis skin; furthermore, head movements of the participant induced noise in the signal. The most robust eye tracker was the MVoG: the eye camera is attached to the subjectfis head and head movements produce less error. However, it was not always easy to set up the MVoG to detect the pupilfi?!especially with some participants with slanting or narrow eyes.

We conducted several experiments: a total of 300 recordings by 14 subjects and 109,097 read words. In the large-scale experiment, we tested three inexpensive devices using three different evaluation methods: user and document independent; user independent and document dependent; and user dependent and document independent. We obtained the best results with the user-dependent method since it better fitted the subjectfis behavior. However, as a cold start, the system should be employed in a user- and document-independent way: that will be able to predict the number of read words when no other recordings from the same subject are available. If the same subject continues utilizing the system, the performances can be improved by adopting a user-dependent approach. We also computed the cumulative error for each participant. Indeed, the Wordometer sometimes overestimates or underestimates the number of read words; thus, if we cumulate the estimation among the read documents, the error is reduced. In that case, it is not possible to determine the number of read words for each document, only for the whole set.

Our second experiment analyzed the robustness of the Wordometer against specific types of reading behavior: rereading and skipping. Rereading had almost no impact on the Wordometer since the system is based on saccade and fixation features, which are the same in case of rereading. Skipping is more difficult to deal with since it is similar to quick reading, but the reader does not actually read any words. The Wordometer is not robust to skipping: a preprocessing step should be used to prevent confusion while processing the signal with the Wordometer.

Our future work involves establishing a preprocessing step to filter different reading behaviors, such as skimming and skipping. We then intend to compare Wordometer performance while reading on paper as against

on screen with the MVoG and EoG systems. Another future study involves analyzing the impact of layout and fonts on the Wordometer.

ACKNOWLEDGMENTS

This work is supported in part by the JST CREST Grant Number JPMJCR16E1, and the JSPS KAKENHI Grant Numbers 25240028, 15K12172, and 16K16089.

REFERENCES

[1] C. Gurrin, A. F. Smeaton, and A. R. Doherty, "Lifelogging: Personal big data," *Foundations and trends in information retrieval*, vol. 8, no. 1, pp. 1–125, 2014.

[2] E. K. Choe, N. B. Lee, B. Lee, W. Pratt, and J. A. Kientz, "Understanding quantified-selfers' practices in collecting and exploring personal data," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 1143–1152.

[3] P. M. Hurvitz, A. V. Moudon, B. Kang, B. E. Saelens, and G. E. Duncan, "Emerging technologies for assessing physical activity behaviors in space and time," *Emerging Technologies to Promote and Evaluate Physical Activity*, p. 8, 2014.

[4] K. Kitamura, T. Yamasaki, and K. Aizawa, "Food log by analyzing food images," in *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 999–1000.

[5] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 477–486.

[6] S. Abdullah, E. L. Murnane, M. Matthews, M. Kay, J. A. Kientz, G. Gay, and T. Choudhury, "Cognitive rhythms: unobtrusive and continuous sensing of alertness using a mobile phone," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 178–189.

[7] P. T. Terenzini, L. Springer, E. T. Pascarella, and A. Nora, "Influences affecting the development of students' critical thinking skills," *Research in higher education*, vol. 36, no. 1, pp. 23–39, 1995.

[8] O. Augereau, K. Kise, and K. Hoshika, "A proposal of a document image reading-life log based on document image retrieval and eyetracking," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 246–250.

[9] K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise, "The wordometer–estimating the number of words read using document image retrieval and mobile eye tracking," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 25–29.

[10] S. Ishimaru, K. Kunze, K. Kise, and A. Dengel, "The wordometer 2.0: estimating the number of words you read in real life using commercial eog glasses," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 293–296.

[11] K. Rayner, T. J. Slattery, and N. N. Bélanger, "Eye movements, the perceptual span, and reading speed," *Psychonomic bulletin & review*, vol. 17, no. 6, pp. 834–839, 2010.

[12] M. Shelhamer and D. C. Roberts, "Chapter 6 - magnetic scleral search coil," in *Vertigo and Imbalance: Clinical Neurophysiology of the Vestibular System*, ser. Handbook of Clinical Neurophysiology. Elsevier, 2010, vol. 9, pp. 80 – 87. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1567423110090064

[13] R. Barea, L. Boquete, S. Ortega, E. López, and J. Rodríguez-Ascariz, "Eog-based eye movements codification for human computer interaction," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2677–2683, 2012.

[14] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, "Eye movement analysis for activity recognition using electrooculography," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 741–753, 2011.

[15] C. Galdi, M. Nappi, D. Riccio, and H. Wechsler, "Eye movement analysis for human authentication: Critical survey," *Pattern Recognition Letters*, 2016.

[16] K. Kunze, K. Masai, M. Inami, Ö. Sacakli, M. Liwicki, A. Dengel, S. Ishimaru, and K. Kise, "Quantifying reading habits: counting how many words you read," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 87–96.

[17] Y. Shiga, T. Toyama, Y. Utsumi, K. Kise, and A. Dengel, "Daily activity recognition combining gaze motion and visual features," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 2014, pp. 1103–1111.

[18] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato, "Coupling eye-motion and ego-motion features for first-person activity recognition," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 1–7.

[19] A. I. Adiba, N. Tanaka, and J. Miyake, "An adjustable gaze tracking system and its application for automatic discrimination of interest objects," *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 2, pp. 973–979, 2016.

[20] C. Holland and O. V. Komogortsev, "Biometric identification via eye movement scanpaths in reading," in *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, 2011, pp. 1–8.
[21] K. Rayner, K. H. Chace, T. J. Slattery, and J. Ashby, "Eye movements as reflections of comprehension processes in reading," *Scientific Studies of Reading*, vol. 10, no. 3, pp. 241–255, 2006.
[22] O. Augereau, H. Fujiyoshi, K. Kunze, and K. Kise, "Estimation of english skill with a mobile eye tracker," in *Adjunct Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2016 ACM International Symposium on Wearable Computers*. ACM, 2016, pp. 1777–1781.
[23] O. Augereau, H. Fujiyoshi, and K. Kise, "Towards an automated estimation of english skill via toeic score based on reading analysis," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 1285–1290.
[24] R. Biedert, A. Dengel, M. Elshamy, and G. Buscher, "Towards robust gaze-based objective quality measures for text," in *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012, pp. 201–204.
[25] K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling, "I know what you are reading: recognition of document types using mobile eye tracking," in *Proceedings of the 17th annual international symposium on International symposium on wearable computers*. ACM, 2013, pp. 113–116.
[26] M. Penttinen and E. Huovinen, "The early development of sight-reading skills in adulthood a study of eye movements," *Journal of Research in Music Education*, vol. 59, no. 2, pp. 196–220, 2011.
[27] H. R. Gudmundsdottir, "Advances in music-reading research," *Music Education Research*, vol. 12, no. 4, pp. 331–338, 2010.
[28] C. Rigaud, T.-N. Le, J.-C. Burie, J.-M. Ogier, S. Ishimaru, M. Iwata, and K. Kise, "Semi-automatic text and graphics extraction of manga using eye tracking information," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 2016, pp. 120–125.
[29] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological bulletin*, vol. 124, no. 3, pp. 372–422, 1998.
[30] G. Buscher, A. Dengel, and L. van Elst, "Eye movements as implicit relevance feedback," in *CHI'08 extended abstracts on Human factors in computing systems*. ACM, 2008, pp. 2991–2996.
[31] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 2000, pp. 71–78.
[32] M. A. Case, H. A. Burwick, K. G. Volpp, and M. S. Patel, "Accuracy of smartphone applications and wearable devices for tracking physical activity data," *Jama*, vol. 313, no. 6, pp. 625–626, 2015.
[33] C. Lennon and H. Burdick, "The lexile framework as an approach for reading measurement and success," *electronic publication on www. lexile. com*, 2004.
[34] A. Gibaldi, M. Vanegas, P. J. Bex, and G. Maiello, "Evaluation of the tobii eyex eye tracking controller and matlab toolkit for research," *Behavior Research Methods*, pp. 1–24, 2016.
[35] R. Biedert, J. Hees, A. Dengel, and G. Buscher, "A robust realtime reading-skimming classifier," in *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, 2012, pp. 123–130.
[36] F. Ehrler, C. Weber, and C. Lovis, "Influence of pedometer position on pedometer accuracy at various walking speeds: A comparative study," *Journal of medical Internet research*, vol. 18, no. 10, 2016.

Received May 2017; revised August 2017; accepted October 2017