Inferring word relevance from eye-movements of readers

Tomasz D. Loboda, Peter Brusilovsky, and Jörg Brunstein
School of Information Sciences
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260

ABSTRACT
Reading is one of the most important skills in today's society. The ubiquity of this activity has naturally affected many information systems; the only goal of some is the presentation of textual information. One concrete task often performed on a computer and involving reading is finding relevant parts of text. In the current study, we investigated if word-level relevance, defined as a binary measure of an individual word being congruent with the reader's current informational needs, could be inferred given only the text and eye movements of readers. We found that the number of fixations, first-pass fixations, and the total viewing time can be used to predict the relevance of sentence-terminal words. In light of what is known about eye movements of readers, knowing which sentence-terminal words are relevant can help in an unobtrusive identification of relevant sentences.

Author Keywords
Reading, eye movements, eye tracking, text relevance, word-level relevance, implicit indicator, information seeking, user study.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Evaluation/methodology, Input devices and strategies*; H.1.2 Models and principles: User/Machine Systems—*Human factors*

INTRODUCTION

Computer software can be called "intelligent" for many reasons. One of them is being able to tell what are the current informational needs of the user, and thus, which documents (e.g., books, scientific articles, Web pages) or parts of documents (e.g., sections, paragraphs, sentences, words) are relevant to them. Establishing that relevance has been a task undertaken by a number of information systems.

In this article, we investigate the possibility of using eye movements to infer which words are relevant to readers who are engaged in an information seeking task. This research has implications for tasks such as automatic text summarization and user modeling. Because of the ubiquity of the written word, results we present here could have an impact on a broad spectrum of information systems. We base our analyses on over 30 years of research on eye movements in reading.

We deal with the notion of *relevance* which we view as a characteristic of a resource (in our case, a word) that makes it satisfy the user's current informational needs or interests, or helps them to achieve their current goals. In that respect, relevance is the manifestation of informational needs, interests, or goals, irrespective of which of these underlying factors is driving it.

Examples of software the goal of which is to infer informational needs of its users include proactive search agents, recommender systems, and information retrieval systems. A proactive search agent resides on the user's computer and monitors search queries they submit and the way they interact with search results. It uses that information to make additional queries on the user's behalf and retrieve new and potentially relevant documents [2, 37]. A recommender system watches for indicators of the user's interests while they browse through a collection of items. It uses that information to maintain a profile of the user and recommend other potentially useful items [3, 8]. An information retrieval system responds to the user's query with a list of results. It then monitors how the user interacts with that list (e.g., which documents they access) and uses that information as relevance feedback to improve on the original results or to personalize their order [13, 36].

While many of these applications and systems use explicit ratings to model informational needs, one of the main challenges they have faced has been to deduce these needs through observation rather than inquiry. Consequently, over the last 10 years a number of projects focused on unobtrusive elicitation of user interests and needs. Many research projects explored a range of relevance indicators (e.g., time spent reading, downloading, printing, etc.) which can provide evidence that a document as a whole is of interest to the user [5, 9]. Other projects explored approaches to extract relevance on the level of parts of a document [4, 7, 11].

Eye tracking emerged as a powerful source of information that can be used to achieve both of the above goals, i.e., help to identify relevant documents and locate relevant parts within them [1, 4, 26, 32, 33].

CHI 2009, April 3 - 9, 2009, Boston, MA, USA.

Our work continues to harvest eye movement data in search of relevance markers, but addresses this problem from the perspective of reading. We use what is known about eye movement behavior of readers to infer the relevance of individual words. Our findings suggest that only eye movements made on sentence-terminal words seem to be promising in that capacity. Because the meaning of a sentence is integrated into the situation model at the sentence-terminal words (sentence wrap-up effect) [17, 29], our results further suggest that by looking at these words we can hope to identify relevant sentences. The fact that we root our investigation in the research on reading allows us to focus on aspects that have not been inspected yet in the context of inferring relevance of text documents.

In the reminder of this article, we first provide background on implicit relevance indicators and the basics of eye movements in reading. Then, we describe the experiment we ran, discuss the validity of data we collected, and present results of our analyses. After that, we provide a general discussion. We finish by demonstrating a way to operationalize our results.

IMPLICIT RELEVANCE INDICATORS

Users of early software were required to communicate their informational needs explicitly. However, it quickly became obvious that the volume of explicit feedback was insufficient to reliably establish those needs. These observations motivated research on implicit relevance indicators. To date, the potential of several indicators have been investigated.

The first group of indicators that have been looked at were document-level actions that users perform for themselves. For example, a number of authors suggested that time spent reading can be an indicator of interest [5, 35]. Claypool provided an early evaluation of this assumption demonstrating a positive correlation between time spent reading and interest [5]. Other researchers explored the use of actions which demonstrate larger commitment such as bookmarking, downloading, and printing. In particular, Kim demonstrated that in the news domain, printing can be a more reliable interest indicator than time spent reading [19]. More recently, tagging emerged as a promising sign of user interest. Several research endeavors demonstrated that the use of tags for modeling interest can improve the precision of recommendations [35, 41]. Tagging also has an added benefit of showing aspects in which the document is of interest to the user [22]. For example, all user tags may be related to music.

Once a document is accessed, we can infer interest by monitoring how the user interacts with it. Actions such as scrolling, mousing-over, and following links have been shown as promising in that capacity [5, 7, 9, 11, 19, 35]. For example, Claypool found a positive correlation between scrolling and interest [5]. Hijikata demonstrated that tracking the mouse motion can help in identifying text fragments which the user might be interested in and thus improve the precision of the user's profile [11].

A combination of document-level and within-document indicators has also been explored. For example, the WAIR system used bookmarking, scrolling, time spent reading or off task, and clicking [35]. By learning the impact of each factor, WAIR was able to achieve a filtering performance superior to that of traditional document-level relevance feedback.

Overall however, the results of research on implicit interests indicators are mixed. While a number of indicators have been reported as useful, no "silver bullet" has been discovered. The effectiveness of many indicators was frequently only moderate and context-dependent, i.e., the same indicator has been shown to work well in one context but to be virtually useless in another. For example, time spent reading has been reported as both a good [5] and a bad [19] indicator.

Finally, a relatively new research direction on information access systems has used eye movement data as a source of implicit relevance [32, 33] and to aid generation of implicit search queries for proactive search systems [1, 26]. Eye tracking technology has also been employed to identify relevant parts of documents that can be used as a source of relevance feedback in personalized search systems [4]. The current research expands on these prior efforts by viewing a text document as being more fine grain.

EYE MOVEMENTS IN READING

Because only about 2° of our vision field can be subjected to foveal processing, our eyes are in constant motion. When viewing a static stimulus, for example a page of text or a picture, our eyes do not move smoothly though. Instead, they make rapid movements called *saccades*. The purpose of a saccade is to bring a new part of the stimulus into fovea. Information from the stimulus is not extracted during saccades, when the eye moves at very high velocities, but only between them, when the eye remains relatively still making so-called *fixation*.

Most saccades made by readers advance the eye forward. Occasionally though, the eye makes a regressive saccade (or simply *regression*) to an earlier part of text. Regressions are often induced by comprehension problems, but are also believed to happen due to oculomotor errors (after a forward saccade overshoots the target) [31].

Many factors influence how likely a word is to be fixated and for how long it will be fixated. These factors include the word's length, its frequency in the language (as determined from corpus data), how early in live it is acquired, how familiar it is, its position in the sentence, the context it appears in, or even the quality of the print and the line length [28]. For instance, short and high-frequency words are more likely to be skipped by readers than are long and low-frequency words.

Certain information is acquired from the left and the right of fixations. For readers of left-to-right alphabetical orthographies (such as English), word length information is acquired from as far as 14-15 character spaces to the right of fixation

(so called *perceptual span*) [23]. However, a word can be identified typically at 7-8 character spaces to the right (so called *word identification span*) [28].

There are various ways of analyzing eye movement reading data. Global averages can be calculated for larger segments of text such as paragraphs. These measures have been shown to reflect reading difficulty (manifested by, e.g., the increase in the average fixation duration). Although these global aggregates may be useful, a more fine grain measures are needed in order to understand cognitive processes on a moment-to-moment basis [30]. In this article, we focus on word-based eye movement variables used pervasively in the reading literature.

METHOD

Subjects

Fourteen students with normal or corrected to normal vision participated for monetary compensation. Seven of them were males and the mean age was 26 years (SD = 9). In this article, we discuss the analysis of eye movement data from six of these subjects. We analyze a subset of data because due to the experimenter's error some subjects were presented with text shown in very small font.

Apparatus

A Tobii 1750 eye tracker was used to display the stimulus at a resolution of 1024x768 pixels and to record the eye movements of the subjects. The eye tracker has a spacial resolution of 0.25 degree and a 50 Hz sampling rate. The subjects viewed the text binocularly at a distance of about 60 cm; 3 characters of text equaled approximately 1° of visual angle. The experiment took place in a usability laboratory with a constant ambient light. No head stabilization was used.

Materials

The materials consisted of a collection of 20 news reports on the December 2002 Texas prison break. Each report was no longer than one page. We excluded the first report because eye movements the subjects made while reading it got mixed with eye movements they made on the demographic questionnaire immediately preceding that news report. We analyzed eye movements made on the remaining 19 reports. These reports had an average of 20.5 lines of text.

Procedure

The experiment lasted approximately 1 hour. The subjects were told to read the news reports and mark those parts of the text that answered the question of "how dangerous the seven escaped convicts were." After completing a nine-point eye tracker calibration routine, they filled out an on-screen demographics questionnaire and read all 20 news reports.

Each news report was presented twice. During the first presentation, the subjects read the report only. The second presentation, which followed immediately, allowed them to mark relevant parts of text. We hoped that this design will reasonably well separate eye movements characteristic to reading and annotation. To make sure the subjects read the text, they were interrupted eight times between the reports and asked to type in an on-screen form how much the report they just read added to their understanding of how dangerous the escapees were. When annotating, the subjects were free to mark text as they felt necessary, i.e., individual words, phrases, or whole sentences. In this article, we analyze only the eye movements recorded during reading (not annotation).

The Data Set

In the analyses we report in this article, we treated the word as the unit of analysis. In order to calculate word-level eye movement variables, we needed to associate fixations with individual words. In doing so, we adhered to the following rules: (1) blank spaces mark word boundaries, (2) blank spaces preceding a word are part of that word, and (2) punctuation characters immediately preceding and following a word are part of that word. For example, all fixations made on the string " *"welcome,*" (i.e., the sequence: space, quote, word, comma) would be treated as belonging to the word *welcome* (see also Figure 2). When calculating the length of words we trimmed both spaces and punctuation characters. For example, the length of the word " *"welcome,*" would be seven. Note that almost all reading studies use the number of letter spaces as the length metric [31].

To account for the perceptual span and word identification span, we focused on words that were *seen* as opposed to only those that were fixated. We considered a word as seen if it was either fixated or immediately followed a fixated word, irrespective of the length of either of the two words. By looking at seen words we were able to calculate the number of words skipped.

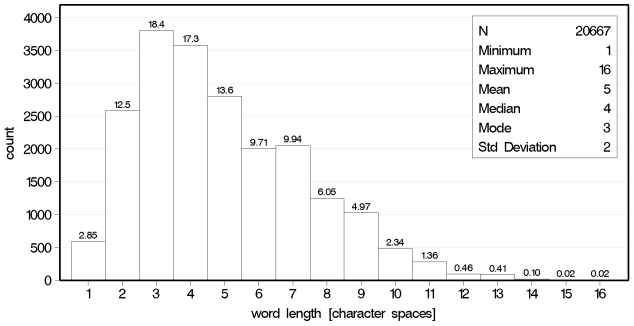

Figure 1. Histogram of the length of words seen by the subjects. Percentages are printed on top of the bars.

Figure 1 shows the distribution of length of all words that were seen by the subjects. Because the number of words longer than 11 character spaces was small (210 words; 1%), we excluded them to avoid unreliable estimates. Overall, we analyzed data consisting of 20,457 words (per subject: M = 3424.67, SD = 219.79). Of them, a total of 2659 (per subject: M = 443.17, SD = 271.53; 13%) were marked as relevant by the subjects. The only other type of words we excluded were numbers.

We generated all data using the Pegasus software [21]. We excluded fixations shorter than 80 ms and longer than 1200 ms as is done commonly in reading studies.

Analyses and Measures

We run two analyses. The first analysis aimed at checking if the eye movements we recorded were indigenous to reading thus validating or invalidating our data set. The second analysis addressed the main goal of this research, i.e., investigating if values of word-based eye movement metrics are contingent upon the word being relevant or non-relevant. In these analyses, we focused on variables that capture both first-pass reading[1] and rereading (compare, e.g., GD and TT below). These variables can be divided into two categories:

- Inspection counts:
  - Number of fixations (NF)
  - Number of fixations excluding skipped words (NF.NZ)
  - Number of first-pass fixations (NFPF)
  - Number of first-pass fixations excluding skipped words (NFPF.NZ)
  - Number of regressions (NR)
  - Number of first-pass regressions (NFPR)
- Inspection durations:
  - Single fixation duration (SFD; defined only for words fixated exactly once)
  - First fixation duration (FFD)
  - Gaze duration (GD; the sum of all first-pass fixations)
  - Total viewing time (TT; the sum of all fixations)

Note that inspection probabilities are usually reported in the reading literature. The reason why we focus on counts is that they are easier to apply in practice, e.g., to construct an algorithm for inferring word relevance from eye movements of readers. It is for the same reason that we include the non-zero (NZ) versions of the NF and NFPF measures. These NZ measures will always yield larger estimates of fixation number per word because skipped words (i.e., those with zero fixations) will not be included in the analysis. All inspection durations we report in this article are given in milliseconds.

Eye movement measurement generates observations that are both correlated and non-normal. That is why conventional methods for inferential data analysis (such as analysis of variance and general linear regression) are not applicable. To properly address these aspects of our data we used generalized linear mixed models (GLMMs). We obtained all result using SAS System version 9.2 [34] and report exact two-tailed p-values.

DATA VALIDITY

Before commencing with the word-relevance analysis, we wanted to make sure that the eye movements we recorded were characteristic of reading. If that was not the case, that would mean that the subjects approached looking for relevant parts of text by means of strategies different than reading, for example skimming. In this section, we try to establish if the reading scenario is plausible and if the signal present in our data is strong enough to permit correct word-relevance analysis results.

A typical reading speed is around 250 wpm (words per minute) [28]. The subjects in our experiment read at a rate of about 190 wpm. It is possible that this lower rate was due to the fact that they read more carefully to assess the relevance of the text. Indeed, the presence of reading perspective (e.g., personal interest in the topic) has been found to render reading times longer [18]. When asked to skim the text, normal readers produce rates of 600-700 wpm [28]. That is much higher than what we observed and therefore skimming seems implausible.

Proficient readers of English skip about a third of the words [38]. Consistent with that, our subjects skipped 26%. Additionally, 10-15% of fixations made by readers are regressions [27]. Consistent with that, in our experiment, an average of 20% of all fixations the subjects made were regressions (14% during first-pass reading).

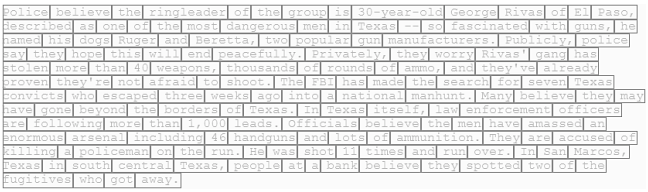

Figure 2. Regions of interest for the shortest news story presented in the experiment (the figure has been reduced; the interface of the Web browser used for presentation is not shown).

While it is comforting that the above aggregate measures support the reading scenario, they provide a very superficial insight. They do not provide any indication of recording accuracy. The eye tracker we used does not provide the state-of-the-art accuracy. Its user's manual explicitly suggests using buffer zones of 30 pixels to separate adjacent regions of interest. Such large buffer zones were impractical because our stimulus was text. We defined our regions of interest, one per word, to touch both horizontally and vertically (Figure 2). This way we minimized the number of off-text fixations, which would be discarded because they say nothing about the relevance of the text.

We were certain that some fixations were recorded incorrectly as belonging to a word different than the one actually fixated. To make sure that this contamination did not introduce too much noise, we attempted to replicate four of the most robust pre-lexical and lexical word-level effects: length [12, 25, **?**], frequency [12, 15, 16, 27], age of acquisition [14, 15], and familiarity [14, 39]. We calculated the length of words as described earlier in this section and obtained the values of the other three variables from the MRC Psycholinguistic Database (as shown on histograms from Figure 3, not all words in our data set had all of these ratings).

Word-Level Effects

Most evidence supporting the existence of word-level effects comes from sentence reading studies. In our experiment, the

[1]First-pass reading starts when the word is fixated for the very first time and ends when the eye moves to another word.

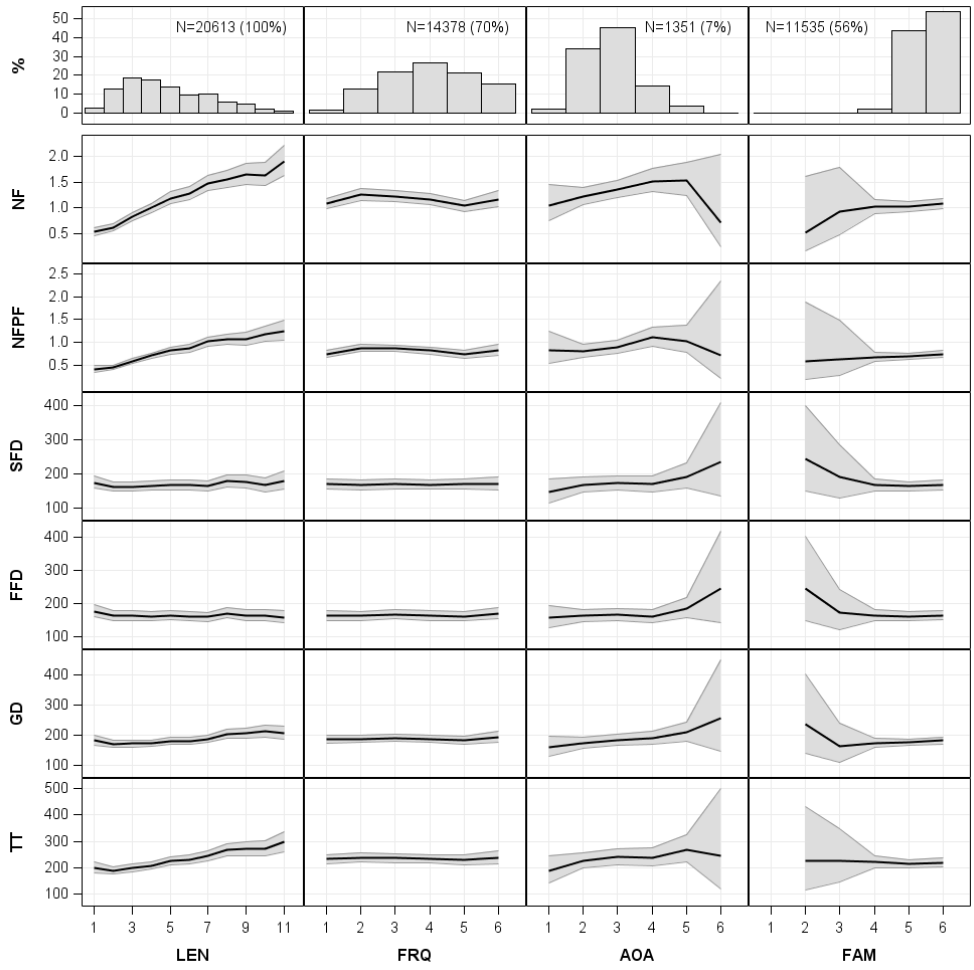

Figure 3. Histograms and effects ($M \pm CI_{95\%}$) of word length (LEN; number of letters), word frequency (FRQ; Kucera-Francis written norms, log_{10}), word age of acquisition (AOA; higher means learned later in life), and word familiarity (FAM; written norms; higher is more familiar). The effects of LEN and FRQ were estimated jointly. When estimating the effects of AOA and FAM, we controlled for FRQ. N shows the number of words having a particular rating, e.g., only about 7% of words had AOA ratings.

subjects read longer passages of text. While reading studies employing larger portions of text exist, the task they employ is usually framed as natural reading or reading for comprehension. In our experiment, the task was to search for relevant parts. For these reasons, some deviations from the results reported by "pure" reading studies would not be unexpected. However, the absence of all of them would, in our opinion, invalidate the data.

We investigated the effect that word length, frequency, age of acquisition, and familiarity had on the following eye movement variables: NF, NFPF, SFD, FFD, GD, and TT. We did that by fitting a set of GLMMs to the entire data set (no separation into relevant and non-relevant words at this stage). Because word length and frequency are invariably confounded in natural language, we estimated their effects jointly. When estimating the effects of age of acquisition and familiarity we controlled for word frequency.

Prior effects plots are shown in Figure 3. The word length effect (the first column) can be clearly seen on four of the plots. Longer words attracted more fixations (NF) and first-pass fixations (NFPF), and were processed for longer during first-pass reading (GD) and overall (TT). The two remaining fixation duration measures (SFD and FFD) were insensitive to word length.

Word frequency (the second column) affected only the number of fixations (NF) and first-pass fixations (NFPF), and in a very subtle way. None of the inspection duration measures turned out to be sensitive to word frequency. This is inconsistent with reading scenario. Usually, the frequency effect is noticeable even when length is controlled for [6]. However, frequency effects have been found to disappear altogether when subjects are asked to search for a target word [27]. Therefore, it is possible that when the task is framed as reading to find answers to a concrete question, eye movement patterns start resembling a mix of reading and target word search. Of course, this line of reasoning implies that readers

know which words to look for. In the case of our study, these were parts of text talking about the danger that the seven convicts posed. The subjects could have expected words such as "weapon," "gun," "knife," "kill," etc, especially after having read the first few pages. Another piece of relevant evidence is that the frequency effect is attenuated as words are repeated. More specifically, after the third encounter of a word in a short passage, low- and high-frequency words become indistinguishable with respect to local eye movement measures [6]. This phenomenon could also help to explain the lack of frequency effect in our data. Finally, it might simply be that once the variation due to word length was accounted for, there was simply too much noise present in the data for the frequency effect to be noticeable.

The effect of word age of acquisition (the third column) is visible on all six graphs. The number of fixations (NF) and first-pass fixations (NFPF) increased slightly with the age of acquisition, but dropped for the last class (AOA=6; words acquired the latest in life). Despite that, all first-pass reading inspection durations (SFD, FFD, and GD) increased for words we learn later in life.

The influence of word familiarity (the fourth column) is visible on five graphs. The least familiar words (FAM=2) received fewer fixations (NF) and first-pass fixations (NFPF) than the more familiar ones, but they were processed longer, as evidenced by the first-pass reading duration measures (SFD, FFD, and GD). The total viewing time (TT) was unaffected by word familiarity. Note that data points for the first class (FAM=1) are missing in the plots because that class had too few words.

To summarize, we found word length, age of acquisition, and familiarity effects in our data. Despite the fact that we did not find frequency effects we think that the trends in the remaining three variables provide enough evidence for the validity of this data set, especially given that the task our subjects faced was not framed as natural reading or reading for comprehension. Before we move on to word relevance however, there is one additional comment we want to make.

Figure 3 also shows the distributions of the four above psycholinguistic variables and the proportion of words with each of the three ratings. About 70% of words the subjects saw had frequency ratings, about 7% had age of acquisition ratings, and about 56% had familiarity ratings. Because of the large number of missing values for the two last variables, we do not look at them in the section devoted to word relevance.

WORD RELEVANCE EFFECT

To understand if and how word relevance may affect eye movements of readers we compared two disjoint sets of relevant and non-relevant words defined by their position in the sentence. The first set consisted of sentence-terminal words. The reason to treat sentence-terminal words differently is that they tend to be read for a longer time than sentence-internal words. Readers pause for longer at these words to understand the meaning of the sentence and how it relates to what they have read so far. This elevation in reading time is known as the *sentence wrap-up* effect and it is one of the manifestations of integrative processes in reading [16, 29]. Therefore, by looking at sentence-terminal words only we were able to check if eye movement variables could help to identify relevant sentences.

The second set consisted of sentence-internal words. Looking at these words allowed us to check if eye movement variables could help to identify relevant words irrespective of their syntactic role. If that was the case, then we could hope to identify arbitrary chains of relevant words in the middle of a sentence.

Models we fitted were given as

$$\log(y_{ij}) = \mu + \lambda_{ij} + \gamma_{ij} + \rho_{ij} + \pi_{ij} + (\rho\pi)_{ij} + s_j + e_{ij}, \quad (1)$$

where y is the response variable, μ is the overall mean, λ is the word length (11 levels), γ is the word frequency (6 levels; Kucera-Francis written norms, log_{10}), ρ is the word relevance (binary), π is the word position (binary), s is the subject random effect, e is the random measurement error, i indexes words and j subjects, $e_{ij} \sim iid\ \mathrm{N}(0, \sigma_e^2)$, and $s_j \sim iid\ \mathrm{N}(0, \sigma_s^2)$.

We assumed the distribution $y_{ij}|s_j$ to be Poisson for inspection counts and LogNormal for inspection durations (Figure 4; see also [10]). To monitor for dispersion problems in count models we looked at the variance of Pearson residuals [20] and, if necessary, allowed the estimation of an additional dispersion parameter.

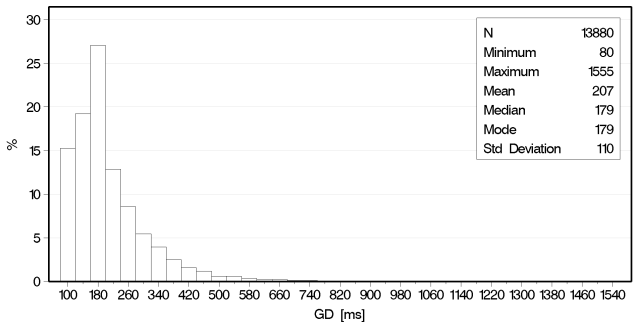

Figure 4. Histograms of gaze duration (GD). Distributions of the other three inspection durations had similar profiles.

The reason why it is important to keep the word length (λ) and word frequency (γ) covariates in the model is that these two effects are among the most important word-level effects that influence where and when we move our eyes during reading. They should be accounted for before any remaining variation in the response variable can be attributed to the quantities of interest[2]. For succinctness, we do not report p-values for these two terms even though they explained a significant amount of variation in all of the models we fitted.

The results are shown in Table 1. Relevant sentence-terminal words received more fixations (NF and NF.NZ) and first-

[2]We report results (p-values and CIs) of type 3 analyses

Table 1. Word relevance analysis results. None of the NR, NFPR, SFD, or FFD models was significant (all $p \geq .215$). Word length (λ) and frequency (γ) were significant in all models but are omitted here for brevity.

							Non-relevant		Relevant	
y	x	p	F	π	p	t	$\boldsymbol{\mu}$	95% CI	$\boldsymbol{\mu}$	95% CI
NF	ρ	.237	1.40							
	π	$.002^{**}$	9.62							
	$\rho \times \pi$	$.011^{**}$	6.46	s-int	$.038^{*}$	2.08	**1.16**	$1.06 - 1.26$	**1.09**	$.99 - 1.21$
				s-term	$.050^{*}$	1.96	**1.19**	$1.07 - 1.32$	**1.39**	$1.17 - 1.64$
NF.NZ	ρ	.059	3.56							
	π	$.003^{**}$	8.68							
	$\rho \times \pi$	$.004^{**}$	8.43	s-int	.123	1.54	**1.57**	$1.47 - 1.70$	**1.66**	$1.53 - 1.79$
				s-term	$.012^{**}$	2.53	**1.58**	$1.47 - 1.70$	**1.80**	$1.61 - 2.01$
NFPF	ρ	.140	2.18							
	π	$.002^{**}$	10.04							
	$\rho \times \pi$	$.016^{*}$	5.76	s-int	.151	1.44	**.81**	$.75 - .87$	**.78**	$.72 - .85$
				s-term	$.041^{*}$	2.04	**.83**	$.76 - .91$	**.96**	$.83 - 1.11$
NFPF.NZ	ρ	$.022^{*}$	5.27							
	π	$.009^{**}$	6.89							
	$\rho \times \pi$	$.015^{*}$	5.91	s-int	.852	1.20	**1.10**	$1.04 - 1.16$	**1.10**	$1.04 - 1.18$
				s-term	$.013^{**}$	2.49	**1.11**	$1.04 - 1.18$	**1.24**	$1.12 - 1.36$
GD	ρ	.146	2.12							
	π	$.039^{*}$	4.27							
	$\rho \times \pi$	.128	2.32	s-int	.927	.09	**187.58**	$176.69 - 199.14$	**187.30**	$175.46 - 199.96$
				s-term	.119	1.56	**190.38**	$177.51 - 204.18$	**205.70**	$184.77 - 228.99$
TT	ρ	.053	3.76							
	π	$.028^{*}$	4.85							
	$\rho \times \pi$	$.031^{*}$	4.63	s-int	.734	.34	**234.07**	$218.72 - 250.51$	**232.62**	$215.88 - 250.66$
				s-term	$.032^{*}$	2.14	**234.65**	$216.46 - 254.37$	**266.91**	$234.89 - 303.29$

ρ – word relevance π – word position (s-int: sentence-internal, s-term: sentence-terminal)

pass fixations (NFPF and NFPF.NZ) than did non-relevant words. However, only the difference in NFPF.NZ was significant for sentence-internal words. The incidence of both regressions (NR) and first-pass regressions (NFPR) was almost identical for relevant and non-relevant words, irrespective of their position in the sentence.

Total viewing time (TT) was the only inspection duration measure with reliably different value for relevant and non-relevant sentence-terminal words, with relevant words accumulating longer reading times. Gaze durations (GD) were also longer for relevant sentence-terminal words, but not significantly so. The other two inspection durations (FFD and SFD) turned out to be completely unaffected by word relevance. Word relevance did not interact with word position nor was its effect alone significant for sentence-internal words on neither of the inspection duration measures.

Note that Table 1 does not show results for NR, NFPR, SFD, and FFD models, because all terms had $p \geq .215$.

DISCUSSION

Our data suggests that relevant sentence-terminal words attract more fixations (both during first-pass reading and overall) and are fixated for a longer time (overall only, but with a trend in gaze duration) as compared to non-relevant sentence-terminal words. Because we controlled for word length and word frequency, the two most robust word-level effects, it seems likely that these differences in eye movement patterns are indeed due to word relevance.

It is known that words at the end of the sentence captivate the reader's attention for longer and that this is due (at least in part) to integrative processes (sentence wrap-up effect) [17, 29]. In light of the evidence we collected, it looks like integrative processes that happen sentence-finally take longer when the reader stumbles upon a relevant sentence as compared to non-relevant sentence-terminal word. That indicates, that we could hope to identify relevant sentences just by monitoring eye movements that readers make on sentence-terminal words.

However, because we did not control the relevance of sentence-terminal words we cannot say with certainty that it is the relevance of the sentence in its entirety that made the sentence wrap-up effect more pronounced. That is because it is possible, although unlikely in our opinion, that sentence-terminal words alone were very relevant and that is what drove the increase in the sentence wrap-up effect. A sentence reading study with sentence-terminal word relevance control is needed to provide more decisive evidence.

The reason we do not look at clause wrap-up is that it is the less pronounced of the two effects. Note also, that recent evi-

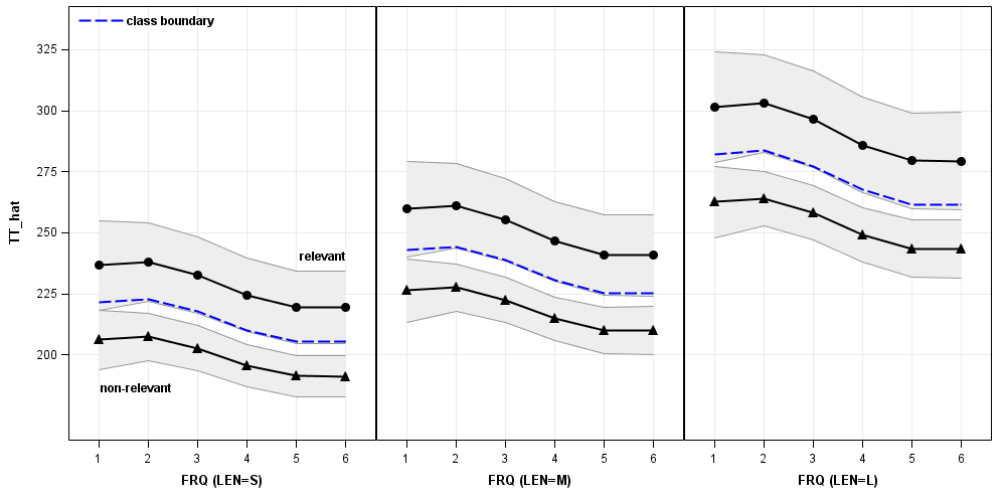

Figure 5. Predicted values ($M \pm SE$) of total viewing time (TT) conditional on word length (LEN) and word frequency (FRQ).

dence suggests that both clause and sentence wrap-up effects are not purely due to the increase in integrative processes, but may additionally be caused by an oculomotor hesitation mechanism [40].

The differences in fixation counts and fixation durations were mostly absent for sentence-internal words. Only the number of first-pass fixations excluding skipped words turned out to be significantly higher for relevant sentence-internal words. It seems then, that we can hope to infer word relevance from eye movements of readers, but only when the words are sentence-terminal. In other words, while it should be possible to identify relevant sentences, arbitrary chains of relevant words are likely to elude identification.

We found it somewhat surprising that there was no difference in the number of regressive saccades made from relevant words, especially the sentence-terminal ones. We expected the subjects to go back to the relevant parts of text more often, especially from the end of the sentence they just read. It appears though, that elevated reading times were enough to determine if words (or sentences) were relevant or not. Therefore, regressions do not seem to be a promising indicator of word relevance.

The fact that we used a remote eye tracker that is known not to provide the best available recording accuracy may be seen as a negative. While this is true, it allowed us to conduct a study in a more ecologically valid setting. Head movements of our subjects were unconstrained (no head stabilization) and therefore the intrusion of the eye movement measurement was minimal. Because of that, our results lend themselves more directly to practical application.

Treating words as units of analysis increased the precision of the estimates we report here by increasing the sample size. Nevertheless, the fact that we analyzed data from only six subjects could rise some concerns. It is true that there is a considerable between-reader variability in fixation durations, saccade lengths, and frequency of regressions. However, the within-subject variability is far greater [28]. Still, it is possible that a group of readers larger than ours would yield results different than the ones we report here. To account for the small number of subject, in our statistical analyses we treated them as a random sample from the universe of all subjects. Finally, eye tracking reading studies involving only several subjects are not unheard of.

In this article, we deal with the notion of word relevance. Relevant words could be seen as interesting. Therefore, genuine interest in a particular topic could yield eye movement patterns similar to those observed when readers are asked to find an answer to a concrete question. That, however, cannot be known prima facie. Because of that, our results may not be directly applicable to the detection or modeling of genuine interest.

The text we used in the current study was a collection of short news reports. Each of these reports was at most a page long. It might be that a more coherent and longer text would result in different patterns of eye movements.

We are not familiar with any other work that investigated the relationship between eye movement behavior of readers and word-level relevance. That is why more studies aimed at understanding this relationship are necessary.

APPLICATION

In this section, we show how one could operationalize the results we discuss above. Because the mean fixation numbers fall between 0 and 2 (Table 1), making use of them to infer the relevance of sentence-terminal words would not be easy if possible at all. For example, how to tell if the reader made 1.19 or 1.39 of a fixation? Because of that, we focus here on the total viewing time (TT).

Because both word length and frequency influence the value of TT (and all other variables for that matter), we need to

take that into account. The length of words the subjects saw in our study was 1-16 characters spaces (see Figure 1), but we analyzed data from words not longer than 11 character spaces. For the purpose of this section, we grouped these words by length into the following three categories: short (1-3 characters), medium (4-6 characters), and long (7-11 characters). This way, the first category contained most of function words, with content words distributed between the two remaining categories. The reason why this is of significance is that function words (articles, conjunctions, prepositions, and pronouns, e.g., "the" or "of") are skipped more often and fixated for less time than content words (nouns, verbs, and adjectives, e.g., "green" or "guitar") [24, 25].

Figure 5 shows predicted values of TT obtained by fitting model (1). One could use the information shown on that graph to decide whether or not a sentence-terminal word is relevant. Given the length (λ) and frequency (γ) of that word and the observed total time TT, the relevance could be defined as 0 if TT $< m(\lambda, \gamma)$, and 1 otherwise. The function $m(\cdot)$ is give as

$$m(\lambda, \gamma) = \widehat{\mathrm{TT}}_{\lambda,\gamma,r=0} + 0.5(\widehat{\mathrm{TT}}_{\lambda,\gamma,r=1} - \widehat{\mathrm{TT}}_{\lambda,\gamma,r=0})$$

and returns the midpoint between the predicted non-relevant and relevant TTs (the dashed line in Figure 5). In this case, TTs equal to or greater than the midpoint will indicate that the word is relevant.

Irrespective of the method, what we deal with here is a binary classification problem. That is, the task is to determine if a sentence-terminal word is relevant or non-relevant given some of its characteristics (e.g., length and frequency) and values of some eye movement variables (e.g., TT). A construction and evaluation of such a classifier is beyond the scope of this article.

CONCLUSIONS

We investigated eye movement behavior of six readers who read several page-long news reports in order to find relevant parts of text. In this article, we have shown that by monitoring their eye movements alone we can hope to identify relevant sentence-terminal words. Given what is known about eye movements of readers, this could help in identifying relevant sentences. The data we collected points towards three local eye movement measures as promising in that capacity: number of fixations, number of first-pass fixations, and the total viewing time. We found almost no evidence that sentence-internal relevant words could be identified from eye movements.

Eye tracking equipment is becoming a more accurate and affordable and therefore pervasive source of information about users. Because of that, the findings we have presented could find their way into and influence the shape of many advanced user interfaces.

Acknowledgments.
This work was performed in the context of GALE project. Angela Brunstein and Rosta Farzan helped with the experimental design.

REFERENCES

1. Ajanki, A., Hardoon, D., Kaski, S., Puolamäki, K., & Taylor, J. (2009) Can eyes reveal interest? Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction*, 19(4), 307–339.

2. Babaian, T., Grosz, B., & Shieber, S. (2002) A writer's collaborative assistant. Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI), 7–14.

3. Billsus, D. & Pazzani, M.J. (2000) A learning agent for wireless news access. Proceedings of International Conference on Intelligent User Interfaces (IUI), 94–97.

4. Buscher, G., Dengel, A., & van Elst, L. (2008) Query expansion using gaze-based feedback on the subdocument level. Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval, 387–394.

5. Claypool, M., Le, P., Wased, M., & Brown, D. (2002) Implicit interest indicators. Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI), 33–40.

6. Clifton, C., Staub, A., & Rayner, K. (2007) Eye movements in reading words and sentences. In *Eye movements: A window on mind and brain* (pp. 341–371). New York: Elsevier.

7. Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005) Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2), 147–168.

8. Fu, X., Budzik, J., & Hammond, K. J. (2000) Mining navigation history for recommendation. Proceedings of International Conference on Intelligent User Interfaces (IUI), 106–112.

9. Goecks, J. & Shavlik, J. (2000) Learning user's interests by unobtrusively observing their normal behavior. Proceedings of International Conference on Intelligent User Interfaces (IUI), 129–132.

10. Harris C.M., Hainline L., Abramov I., Lemerise E., & Camenzuli C. (1988) The distribution of fixation durations in infants and naive adults. *Vision Research*, 28, 419–32.

11. Hijikata, Y. (2004) Implicit user profiling for on demand relevance feedback. Proceedings of the 9th International Conference on Intelligent User Interfaces (IUI), 198-205.

12. Hyönä, J. & Olson, R.K. (1995) Eye movement patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 21, 1430–1340.

13. Joachims, T., Granka, L., Bing Pan, Hembrooke, H., & Gay, G. (2005) Accurately interpreting clickthrough data as implicit feedback. Proceedings of the 28th Annual International ACM SIGIR Conference, 154–161.

14. Juhasz, B.J. & Rayner, K. (2003) Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29, 1312–1318.

15. Juhasz, B.J. & Rayner, K. (2006) The role of age-of-acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition*, 13 (7 & 8), 846-863.

16. Just, M.A. & Carpenter, P.A. (1980) A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–354.

17. Just, M.A., Carpenter, P.A., & Woolley J.D. (1982) Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 3(2), 228–238.

18. Kaakinen, J.K. & Hyona, J. (2005) Perspective effects on expository text comprehension: Evidence from think-aloud protocols, eyetracking, and recall. *Discourse Processes*, 40(3), 239–257.

19. Kim, J., Oard, D., & Romanik, K. (2000) Using implicit feedback for user modeling in internet and intranet searching, Technical Report, College of Library and Information Services, University of Maryland at College Park.

20. Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., & Schabenberger, O. (2006) *SAS for Mixed Models*. 2nd Edition. SAS Publishing, Cary.

21. Loboda, T.D. (2009) Pegasus [computer software]. Pittsburgh, PA.

22. Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006) HT06, tagging paper, taxonomy, Flickr, academic article, to read. Proceedings of the 17th Conference on Hypertext and Hypermedia, 31–40.

23. McConkie, G.W. & Rayner, K. (1976) Asymmetry of the perceptual span in reading. *Bulletin of the Psychonomic Society*, 8, 365–368.

24. O'Regan, J.K. (1979) Eye guidance in reading: Evidence for the linguistic control hypothesis. *Perception & Psychophysics*, 25, 501–509.

25. O'Regan, J.K. (1980) The control of saccade size fixation duration in reading: The limits of linguistic control. *Perception & Psychophysics*, 28, 112–117.

26. Puolamäki, K., Ajanki, A., & Kaski, S. (2008) Learning to learn implicit queries from gaze patterns. Proceedings of the 25th International Conference on Machine Learning, 760–767.

27. Rayner, K. & Raney, G.E. (1996) Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, 3(2), 245–248.

28. Rayner, K. (1998) Eye Movement in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3), 372–422.

29. Rayner, K., Kambe, G., & Duffy, S.A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology*, 53A(4), 1061-1080.

30. Reichle, E.D., Pollatsek, A., Fisher, D.L., & Rayner, K. (1998) Toward a model of eye movement control in reading. *Psychological Review*, 105, 125–157.

31. Reichle, E.D., Rayner, K., & Pollatsek, A. (2003) The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26, 445–476.

32. Salojärvi, J., Kojo, I., Simola, J., & Kaski, S. (2003) Can relevance be inferred from eye movements in information retrieval? Proceedings of Workshop on Self-organizing Maps.

33. Salojärvi, J., Puolamäki, K., and Kaski, S. (2005) Implicit relevance feedback from eye movements. Proceedings of the 15th International Conference on Artificial Neural Networks (ICANN), 513–518.

34. SAS Institute Inc. (2008) SAS 9.2 help and documentation.

35. Seo, Y.-W. & Zhang, B.-T. (2000) A reinforcement learning agent for personalized information filtering. Proceedings of International Conference on Intelligent User Interfaces (IUI), 248–251.

36. Sugiyama, K., Hatano, K., & Yoshikawa, M. (2004) Adaptive Web search based on user profile constructed without any effort from users. Proceedings of the 13th International World Wide Web Conference (WWW), 675–684.

37. Twidale, M., Gruzd, A., & Nichols, D. (2008) Writing in the library: Exploring tighter integration of digital library use with the writing process. *Information Processing and Management*, 44(2), 558–580.

38. Underwood, G. (2005) *Cognitive Processess in Eye Guidance*. Oxford University Press.

39. Williams, R.S. & Morris, R.K. (2004) Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 16, 312–339.

40. Warren, T., White, S.J., & Reichle, E.D. (2009) Investigating the causes of wrap-up effects: Evidence from eye movements and EZ Reader. *Cognition*, 111, 132-137.

41. Zhao, S., Du, N., Nauerz, A., Zhang, X., Yuan, Q., & Fu, R. (2008) Improved recommendation based on collaborative tagging behaviors. Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI), 413–416.