InteRead: An Eye Tracking Dataset of Interrupted Reading

Francesca Zermiani[1], **Prajit Dhar**[2], **Ekta Sood**[1], **Fabian Kögel**,
Andreas Bulling[1], **Maria Wirzberger**[1]
[1]University of Stuttgart, [2]University of Potsdam
[1,2]Germany
{francesca.zermiani, maria.wirzberger}@ife.uni-stuttgart.de
{ekta.sood, andreas.bulling}@vis.uni-stuttgart.de
dhar@uni-potsdam.de, f.koegel@web.de

Abstract

Eye movements during reading offer a window into cognitive processes and language comprehension, but the scarcity of reading data with interruptions – which learners frequently encounter in their everyday learning environments – hampers advances in the development of intelligent learning technologies. We introduce *InteRead* – a novel 50-participant dataset of gaze data recorded during self-paced reading of real-world text. InteRead further offers fine-grained annotations of interruptions interspersed throughout the text as well as resumption lags incurred by these interruptions. Interruptions were triggered automatically once readers reached predefined target words. We validate our dataset by reporting interdisciplinary analyses on different measures of gaze behavior. In line with prior research, our analyses show that the interruptions as well as word length and word frequency effects significantly impact eye movements during reading. We also explore individual differences within our dataset, shedding light on the potential for tailored educational solutions. InteRead is accessible from our datasets web-page: https://www.ife.uni-stuttgart.de/en/llis/research/datasets/.

Keywords: eye tracking, reading, interruptions

1. Introduction

Attention-aware learning technologies (AALT) have been increasingly studied to facilitate learning and engagement in educational contexts. In recent years, the landscape of learning technologies has undergone rapid transformation, accelerated in part by the global upheaval caused by the COVID-19 pandemic, which intensified the demand for digital learning solutions (Kang, 2021). Within this evolving educational landscape, eye tracking has risen to prominence as a powerful methodology for AALT (Hutt et al., 2021). As reading is one of the most prominent aspects of educational contexts and a challenging learning process (Wijekumar et al., 2012), learning technologies for reading have been developed to facilitate teaching, evaluation, and research of reading abilities such as acquisition, comprehension, literacy, and text readability (Litman, 2016; Jacovina and McNamara, 2017; Atun, 2020).

Understanding eye movements during reading can offer unique insights into the lexical processes and cognitive skills associated to reading comprehension and attention mechanisms (Just and Carpenter, 1980; Rayner et al., 1989; Schilling et al., 1998; Engbert et al., 2002). For instance, gaze behavior allows us to measure and model comprehension, attention allocation, individual and disrupted behaviors (Reichle et al., 1998; Rayner et al., 2006, 2010; Reichle et al., 2010). These insights have, in turn, motivated further development of a variety of gaze-based AALT, designed to assist learners' attention during computerized reading (D'Mello, 2019) and optimize the reading process (Sibert et al., 2000; Srivastava et al., 2021). Recently, the integration of larger eye tracking corpora for natural language processing (NLP) purposes has proved beneficial for various tasks, including predicting reader comprehension (Mézière et al., 2023), enhancing text understanding (Barrett and Hollenstein, 2020; Mathias et al., 2020) and interpreting computational language models (Sood et al., 2020).

However, one key challenge remains: a scarcity of real-world gaze behavior data, which encompass interruptions that learners frequently encounter in digital learning environments (Potier Watkins et al., 2020). Existing public corpora largely originate from experiments that are not explicitly designed to capture the disruptions prevalent in educational environments, which can significantly impact focus and memory (Naveh-Benjamin et al., 2007; Armendariz et al., 2021). Additionally, current gaze-based AALT exhibit limited adaptability to individual learner experiences, with intervention strategies supporting only specific learners (Hutt et al., 2021). Despite being still insufficiently explored (Bai et al., 2014), investigating the influence of individual differences on how we handle interruptions while reading holds the potential to enhance the design of adaptive interventions and AALT.

To address these limitations, we introduce *InteRead*[1], the interrupted reading eye tracking corpus. InteRead is a novel eye tracking dataset featuring 50 participants engaged in a self-paced reading task of an excerpt from an English fictional text, deliberately interrupted to simulate naturalistic scenarios. This corpus inherits the advantages of larger eye tracking reading corpora with naturalistic stimuli, offering researchers versatility for investigating specific linguistic processes and exploring the cognitive reading skills inherent in the learning process (Hollenstein et al., 2022). More specifically, InteRead offers the opportunity to further examine the impact of interruptions on reading as well as the influence of individual differences on how we recover from such interruptions. Ultimately, it provides additional insights to refine gaze-based AALT for a more personalized reading experience.

2. Related Work

Gaze-Based Attention-Aware Learning Technologies for Reading Attention-aware learning technologies (AALT) are educational software designed to track, respond to, and model learners' attentional states, especially during disruptions (D'Mello, 2019). Eye tracking has emerged as a valuable tool for AALT development since it captures gaze behavior and allows for inferring learners' attention (Gluck et al., 2000; Conati et al., 2013; D'Mello et al., 2012; Hutt et al., 2016). In educational contexts, reading and understanding text is one of the most complex and challenging cognitive skills to learn and teach (Elleman and Oslund, 2019; Smith et al., 2021). Research showcasing the strong link between eye movements and the cognitive processes involved in reading (Just and Carpenter, 1980) has resulted in the development of various computational models (Reichle et al., 1998; Rayner, 1998) and applications designed to understand and enhance the reading experience (Kunze et al., 2013; Rzayev et al., 2018). Furthermore, gaze-based AALT have been employed to detect disruptions and attentional shifts during digital reading (D'Mello et al., 2016; Mills et al., 2021) and assist readers in re-focusing (Jo et al., 2015; Mariakakis et al., 2015; Srivastava et al., 2021). An open challenge in the development of these systems is the limited capacity to adapt to individual differences, resulting in one-size-fits-all designs (Hutt et al., 2021; Mariakakis et al., 2015).

Interrupted Reading The increasing use of digital devices often leads to interruptions during reading, as they frequently divert our attention to other tasks, prompting a shift from the current activity to a new one (Chevet et al., 2022). Task interruption is defined as a shift of cognitive resources from the most active goal representations and the consequent need for reactivating those representations upon task resumption (Altmann and Trafton, 2002). *Resumption lag time* measures users task performance following an interruption (Trafton et al., 2003). From gaze behavior, previous studies have identified task resumption during reading with increased reading times, higher number and duration of fixations, longer saccades and higher regression frequency (Cane et al., 2012; Cauchard et al., 2012; Chevet et al., 2022). Although individual differences have an impact on how we recover from interruptions, the extent of this is still unclear due to limited data (Werner et al., 2011; Meys and Sanderson, 2013; Bai et al., 2014). This holds especially true for interrupted reading tasks, as there are only a limited number of studies that point at the influence of individual differences on the ability to resume reading effectively (Foroughi et al., 2016; Altamura et al., 2022).

Current Eye Tracking Reading Corpora Currently, there are numerous public eye tracking reading corpora, however they do not account for features present in natural learning environmental. These datasets encompass a range of different purposes to study eye movement behavior in reading. Several English datasets involve adult readers engaging in self-paced reading of multiple short texts (e.g., Frank et al., 2013; Mishra et al., 2016; Luke and Christianson, 2018; Hollenstein et al., 2020; Sood et al., 2021). Similar eye tracking corpora are also available in other languages, such as Hindi (Husain et al., 2015), Persian (Safavi et al., 2016), Chinese (Pan et al., 2022), or in a multilingual setting (Siegelman et al., 2022; Kuperman et al., 2023). Additionally, corpora focusing on specific user groups, such as expert vs. novice readers of scientific texts, monolingual vs. bilingual readers, and various age groups, have shown gaze behavior variations (Kliegl et al., 2004; Cop et al., 2017; Jäger et al., 2021; Siegelman et al., 2022; Kuperman et al., 2023; Yi et al., 2020). While previous educational research has explored gaze behavior and recovery from interruptions during the reading of naturalistic texts (Cane et al., 2012; Jo et al., 2015; Chevet et al., 2022), it is noteworthy that these datasets are not publicly accessible. The absence of publicly available eye tracking datasets for educational research poses a crucial gap, limiting research in the fields of gaze-based AALT and NLP applications for education, particularly those aimed at addressing individual differences in reading. To bridge this gap, we introduce the InteRead dataset, a novel eye tracking

[1]The dataset is accessible from our datasets web-page: https://www.ife.uni-stuttgart.de/en/llis/research/datasets/.

dataset containing interruptions.

3. InteRead Dataset

Our InteRead dataset is a novel publicly available eye tracking reading dataset with interruptions, comprising gaze data collected from 50 participants engaged in an interrupted reading task. (Figure 1). With 5,247 tokens spanning across 28 pages of text, six of which include interruptions, InteRead is designed to further study interrupted reading. It also enables to explore individual differences in reading and resuming from interruptions. Such insights can be subsequently leveraged to foster advancements in adaptive gaze-based AALT and NLP applications to support attention and, ultimately, reading comprehension. Moreover, with the incorporation of linguistic features (see Section 4), InteRead constitutes an additional resource to study the linguistic and cognitive processes underlying reading of a fictional text.

3.1. Participants

We collected data from 57 adult participants (36 female, *M* age = 27.51 years, *SD* = 5.55, range 20–47), who were recruited through internal mailing lists of the University of Stuttgart and social media. Compensation included either a monetary benefit of or a participation certificate for study credits. Participants met specific criteria, including normal or corrected-to-normal eyesight, English proficiency (native speaker, C1, IELTS 6.5+ or equivalent), and the absence of diagnosed attention or reading disorders. The study procedure received approval from the ethics committee of the University of Stuttgart (approval number Az. 22-018). Seven participants were excluded from the final dataset for three specific reasons: (1) having more than 50% missing gaze data points on the pages containing an interruption, (2) triggering fewer than three interruptions out of six, and (3) presenting excessively noisy data on the interruption pages.

3.2. Recording Setup

The data collection was conducted within a controlled laboratory environment with consistent lighting conditions. Participants were seated at a desk equipped with an adjustable head and chin rest, mouse, keyboard, monitor and eye tracker, while the investigator seated at another desk behind a room divider. We used a Tobii Pro Spectrum screen-based eye tracker[2] operating at a sampling frequency of 1200Hz. All stimuli were presented on the native Tobii Pro Spectrum screen (EIZO FlexScan EV2451) with dimensions of 52.8×29.7cm and a resolution of 1920x1080px. The screen was placed 57cm in front of the participants, who positioned their eyes in the center of the Tobii Pro Spectrum headbox. Both the eye tracker and its monitor were connected to a desktop computer running the recording pipeline. For the study, we used PsychoPy and PyGaze (Peirce et al., 2019; Dalmaijer et al., 2014)[3].

3.3. Materials

Our corpus contains a total of 5,247 tokens spanning across 28 pages. Among these, six specific pages (pages 3, 7, 11, 15, 19, 24) were strategically chosen to introduce interruptions intermittently, approximately every three to four pages. The interruption pages remained fixed across participants.

Reading Material The reading material was a selected excerpt from Arthur Conan Doyle's "The Adventure of the Speckled Band" (written in British English and published in 1892). The excerpt consisted of 28 pages with on average 154 words each (*SD* = 22.3, range 94–190 words per page). The text followed the original narrative's structure, for instance, direct dialogues within quotation marks were preserved. Figure 2 shows the stimulus presentation setup. Each page contained 12 lines of text, with the exception of the last page containing seven lines. The text format was set in Courier, a black mono-spaced font at 16px (0.44DVA) character height and 2.5-times line spacing. The text was left set at 500px and vertically centered on a white background.

Interruption Material Interruptions took the form of a light gray dialog box at the center of the screen, while the background remained white and the reading stimulus disappeared. The boxes prompted participants to type a response to one of six distinct opinion questions within a 60-second time frame. The questions pertained to everyday events or topics and were selected from question items employed in prior research (Pashler et al., 2013).

In accordance with prior research (Cane et al., 2012; Jo et al., 2015), the interruptions were triggered by fixation of a target word: 'control' (page 3), 'seemed' (page 7), 'troubled' (page 11), 'died' (page 15), 'opposite' (page 19), 'cross' (page 24). Aligning with established methodologies (Jo et al., 2015; Wirzberger and Rey, 2018), target word selection followed predefined criteria: (i) target words were randomly selected always within the middle line of each page, (ii) the first and last words of the middle line were consistently excluded, (iii) target words were never placed as the first or last

[2]Firmware version 2-6-1.

[3]The interface was implemented using PsychoPy (version 2022.2.3). The eye tracking was managed using PyGaze (version 0.7.4).

Figure 1: Phases of recording trials. Interruptions are triggered during the *pre*interruption phase by fixations reaching a target word. After answering the *interruption* questions within 60 seconds, the reading task resumed. The *resumption lag* phase constitutes the time from interruption offset until normal reading behavior reoccurs, indicating the beginning of the *post*interruption phase.

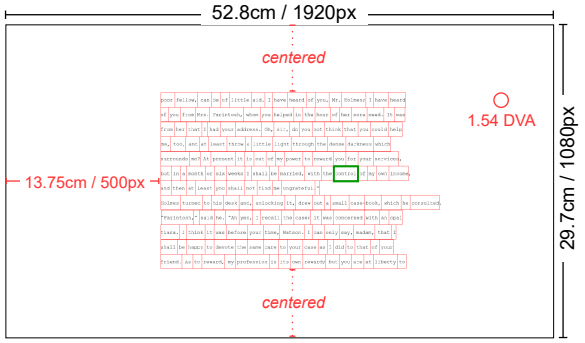

Figure 2: Page setup with bounding boxes and dimensions. Lines covered 1.54DVA. Bounding box for trigger word in green. Annotations in red and green were not presented to the participant.

words of a sentence, ensuring interruptions occurred within sentence contexts, and (iv) both the target words and their corresponding pages remained constant across all participants. Additionally, function words were systematically excluded from the pool of potential target words due to their lower fixation frequency during reading (Rayner and McConkie, 1976; Krejtz et al., 2016).

3.4. Recording Procedure

Participants provided informed consent to share anonymized data with the scientific community. They were informed about potential reading interruptions and instructed to respond to questions about everyday topics when prompted. They were advised to read attentively, as they would be required to complete a brief comprehension test afterward. Prior to the reading task, we also collected demographic data and assessed participants' background knowledge on the content of the text and spatial memory capacities, due to their established interaction with the reading process (Inhoff and Weger, 2005; Smith et al., 2021).

All scores related to the pre- and post-study measures are available in the dataset (see Appendix A and B for detailed descriptions of the pre- and post-test questionnaires).
The eye tracking recording started with a five-point calibration. The participant then read the story – advancing with space key – while the movements of both eyes were captured. There was no option to navigate backward. The space key was restricted on interruption pages, but if the interruption was not triggered on the middle line, it was automatically re-enabled. After reading, participants answered four reading comprehension questions and completed a survey about their interest in the story and how they handled the interruptions.

4. Data Preprocessing

Preprocessing involved two main steps: identifying resumption lag times through manual annotation and extracting features from the raw gaze data. We extracted both eye tracking and linguistic features (see Table 1 for a general statistical account).

Manual Annotation of Resumption Lag Following the postulates of Altmann and Trafton (2002), we define the resumption lag as the time span between the interruption offset and the first stable reading pattern in the pre-interruption text. Resumption lag times, obtained by averaging two human annotators' labels, resulted in six times per participant, one for each interruption. We calculated the Interclass Correlation Coefficient (ICC) (Gamer et al., 2010) to assess annotation reliability[4]. The estimate resulted in a value of ICC = 0.916, above the 0.90 threshold indicating strong reliability (Portney et al., 2009).
The annotators did not participate in data collection and implemented a tool to visualize gaze data

[4]To calculate the ICC, we used the irr package (version 0.84.1) with an average rating of k = 2, absolute-agreement and two-way mixed effects model.

Page(s)	SENT	TOK	TOK LEN	TTR	CON	LOG FREQ	ABS	RT	FIX CNT	NON FIX	FIX DUR	1ST FIX	SAC LEN	REG FREQ
0-2	10	194	3.56	0.46	0.40	12.82	2.75	36.2	367	0.49	204	242	139.71	22.08
3	11	228	3.30	0.52	0.33	13.10	2.64	47.8	401	0.44	205	241	126.74	22.58
4-6	6	187	3.90	0.51	0.41	12.88	2.54	37.1	376	0.47	202	245	142.77	22.57
7	7	193	4.24	0.62	0.43	12.64	2.53	49.0	445	0.34	204	243	129.32	21.94
8-10	7	197	3.81	0.49	0.41	12.74	2.59	36.4	348	0.48	202	244	140.15	21.98
11	9	161	3.50	0.65	0.37	12.71	2.71	30.2	397	0.48	201	234	128.68	21.13
12-14	14	203	3.49	0.44	0.35	12.95	2.65	30.7	311	0.54	201	237	142.70	22.13
15	8	184	3.88	0.64	0.36	12.83	2.74	36.3	380	0.44	199	234	134.42	22.25
16-18	9	187	3.83	0.50	0.37	12.74	2.57	33.2	354	0.50	202	242	137.80	22.23
19	11	198	3.36	0.61	0.37	13.05	2.71	35.2	359	0.47	198	226	127.12	23.60
20-23	10	176	3.58	0.43	0.36	12.89	2.56	33.7	300	0.56	200	236	136.92	22.42
24	6	202	3.66	0.60	0.42	12.80	2.74	42.8	413	0.43	203	237	123.41	22.24
25-27	12	156	3.48	0.47	0.38	13.00	2.74	23.1	319	0.56	202	243	134.01	23.38

Table 1: General statistics of InteRead. Every row alternates between the averaged statistics for a sequence of non-interruption pages and those for a single interruption page (3, 7, 11, 15, 19, 24). SENT: number of sentences; TOK: number of unique tokens; TOK LEN: length of tokens (in chars); TTR: type-token ratio; CON: content words proportion; LOG FREQ: frequency of a token in a natural logarithmic scale; ABS: abstractness/concreteness rating of a token, 1 (purely abstract) to 5 (purely concrete); RT: reading time (in s); FIX CNT: count of fixations for a token; NON FIX: non-fixations proportion; FIX DUR: duration of all the fixations (in ms) for a token; 1ST FIX: fixation duration (in ms) only for the first pass fixation the participant made on a token; SACC LEN: saccade length (in px); REG FREQ: regression frequency. Values for SENT, TOK, FIX CNT, FIX DUR and 1ST FIX are rounded to the nearest integer. Notably, the majority of FIX CNT and FIX DUR hold a value of zero, as highlighted by NON FIX (M = 0.51). Hence, we report the mean values for only the non-zero values for the fixation features: FIX CNT, FIX DUR and 1ST FIX.

as x and y coordinates from interruption offset to page end. They manually selected the start point of the first reading pattern, and the tool converted it to a timestamp, calculating resumption lag time by subtracting the interruption offset timestamp. To ensure the point was from the pre-interruption text, they iterated the annotations.

Gaze Features We observed fluctuations in the timestamps of the raw gaze data obtained through PyGaze and thus resampled the raw data to comply with a strict 1200Hz sampling rate. We linearly interpolated the gaze coordinates between the closest real samples for all valid samples, excluding samples during blinks. We then averaged the coordinates of left and right eyes in the raw gaze data and extracted fixation and saccade events using the REMoDNaV toolkit (Dar et al., 2020)[5].

We extracted the following gaze features: *reading times*, the time (in s) taken by the participant to read a page; *fixation count*, the number of fixations made by a participant for a token; *non-fixations*, a boolean value signifying if a token has fixations; *fixation duration*, the mean duration of all the fixations (in ms) for a token; *first pass fixation*, the fixation duration (in ms) only for the *first pass fixation* the participant made on a token; *gaze duration*, the sum of all first pass fixations on the token; *saccade length*, the mean length of saccades (in px); *regression frequency*, the mean freqeuncy of regressive saccades. We considered regressions are saccades that move backwards in the text. Regression frequency is reported as a proportion of all regressions.

Linguistic Features We parsed the reading material to include linguistic features, such as part-of-speech and dependency relations into our corpus. To parse the text, we used Spacy (Honnibal et al., 2020). We then manually aligned the parsed text to the bounding boxes for each token. Additionally, we also included the abstractness rating and the frequency of the tokens. The abstractness (or concreteness) ratings, taken from Brysbaert et al. (2013), denote the degree to which the meaning of a word is based on a human's perception. The frequency values are extracted from the SUBTLEXus corpus (Brysbaert and New, 2009).

The following linguistic features were extracted: the mean number of *sentences* in a page; the *part-of-speech tags* and *dependency relations* for a token; the total number of *tokens* and *token types*, i.e the number of unique tokens, found in a page; the *token length* (in chars); the *type-token ratio*, the proportion of types to tokens in a given page; the *content words*, a boolean value indicating if a

[5]We applied the default REMoDNaV parameters, except for *pursuit_velthresh = 50000, savgol_length = 0.018, dilate_nan = 0.08*.

token is a content word[6]; the *token frequency in a natural logarithmic scale*; the *abstractness/concreteness* rating of a token, ranging from 1 (purely abstract) to 5 (purely concrete).

5. Data Validation and Analysis

To substantiate the quality of InteRead, we perform various analyses on gaze behavior. As a first step, we focus on interruption pages, delving into how interruptions influence readers' gaze behavior. We subsequently explore how the extracted linguistic features affect eye movements. Ultimately, we delve into individual differences among participants, examining variations in reading and resumption times within our dataset, revealing distinct characteristics of readers.

Interruption Effect on Gaze Behavior To test whether the interruptions have an effect on reading behavior, we follow the approach of Cane et al. (2012). Specifically, we identify the three temporal phases, (1) the pre-interruption phase, (2) the resumption lag time, (3) the post-interruption phase (see Figure 1). Following Cane et al. (2012), we investigate whether the temporal phases have a significant effect on the gaze features that are relevant for higher-level cognitive processes as well as meaning integration – reading time, fixation count, fixation duration, saccade length and regression frequency. We employ a multi variable t-test analysis t-test analysis with a Bonferroni correction and consider the resumption lag times to be part of the postinterrution phase. We observe a significant increase in reading times, fixation counts and saccade lengths following an interruption (see Figure 3). Post hoc analyses reveal that on average readers take significantly longer to read (37s vs. 22s prior to interruption), make more fixations (145 vs. 89) and longer saccades (140px vs. 127px) following an interruption. We find the temporal phase to have no effect on fixation duration, which do not significantly change after an interruption (p=.67 and *p-adjusted*=1.0), and regression frequency (p=.06 and *p-adjusted*=.28).

Word Length and Word Frequency Effect on Gaze Behavior We investigate the influence of logarithmic token frequency and token length on reading gaze behavior. Our experiments focus on first pass fixation and gaze duration as they are associated with lexical aspects like token length and frequency (Schilling et al., 1998; Kliegl et al., 2004; Hollenstein et al., 2022). Prior work in German (Kliegl et al., 2004) shows a significant effect of the logarithmic token frequency and token length on gaze duration, but not on first pass fixation. To discern if such effect is present in our corpus, we use linear mixed effects models with either first pass fixation or gaze duration as dependent variable, and logarithmic token frequency and token length as fixed effects. We take the line, page and participant as random effects (see Appendix C for the full models). Following the approach of Hollenstein et al. (2022), we remove all fixation values under 100ms. We observe that the token length has a non-significant effect on first pass fixations (t = -0.71, p = 0.48), but a significant effect on gaze duration (t = 45.39, p < .001). The logarithmic token frequency instead has a significant effect on both first pass fixations (t = -5.61, p < .001) and gaze duration (t = -2.93, p = .003).

We then test, using one-way ANOVA, the effect of the linguistic features on fixation duration and observe that the part-of-speech tags (F = 3.32) and dependency relations (F = 2.43) for a given token exhibit a significant effect (p < .001). In order to discern which part-of-speech categories and dependency relations show significance, we perform a mixed effects linear regression, with the same random effects as before (line, page and participant). Among the part-of-speech categories, we find adjectives (t = 3.27, p < 0.01) and verbs (t = 2.56, p = 0.01) to be significant. Regarding the dependency relations instead, only appositional modifiers have a significant effect (t = 2.0, p = 0.05) on fixation duration. We then perform a Kendall's rank correlation test to assess the correlation between fixation duration and both word abstractness rating and word length. Both the features show a significant correlation to fixation duration, with p < .001. To determine which features have an effect on whether the participants would skip words during reading, we perform a χ^2 test with Yates' continuity correction. The part-of-speech tags, the dependency relations and the semantic value (content/non-content) of a given word show significant effects on the probability of skipping that word, with p < .001.

Individual Differences in Reading and Resumption Time We initially investigate the distribution of fixation duration across pages and participants, resulting in a mean fixation duration of 202.05ms (SD = 85.68, range = 40–2377.5) and a mean saccade length of 135.783px (3.734cm) (SD = 178.061px (4.897cm) and range 0.050–1690.225px (0.001–46.481cm)). To account for individual differences in reading behavior, we then conduct a two-sample t-test to identify those participants whose reading time significantly varies from the mean reading time of the remaining participants. Among the 50 participants, 32 (18 faster and 14 slower) display reading times significantly different from the mean of the remaining 49 partic-

[6]Content words are words that carry semantic value or meaning and belong to open word classes, such as nouns or verbs.

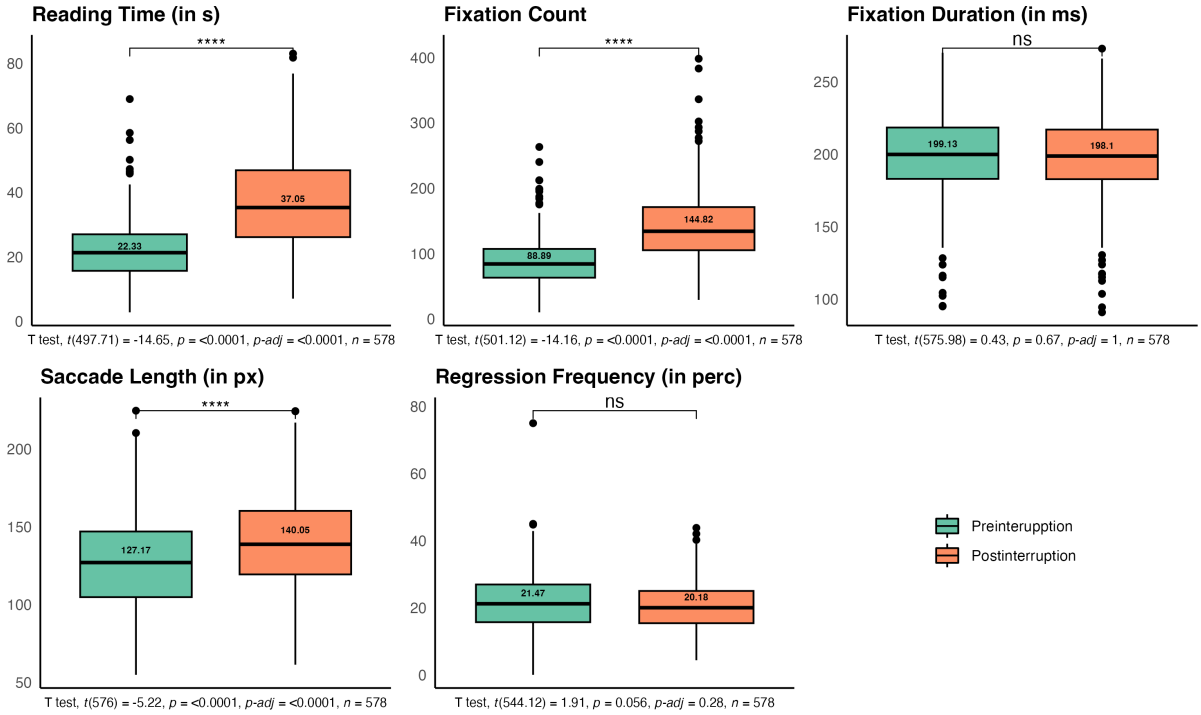

Figure 3: Boxplot showing the distribution of reading time, fixation count, fixation duration, saccade length, regression frequency during pre- and post-interruption phases. The black dots represent the outliers in each distribution. The stars (*) and ns represent if there is a significant difference between the means of two distributions, with the significance denoted by the p-adjusted (p-adj) value. The complete T-test statistics are given below each plot. The values in the boxes are the means of the respective distributions.

ipants (Figure 4). These groups of 18, 18 and 14 participants are clustered into three categories of reading speeds respectively: Fast, Moderate and Slow. Post hoc analyses reveal that slow readers and moderate readers, on average, make respectively 15ms- and 12ms-longer fixations compared to fast readers.

We then test if the created groups have any significant effect on the gaze and linguistic features from InteRead. For these analyses, we use a linear mixed effects model, with page, participant and line as random effects (see Appendix C for the full models). For each of the regression experiments, we determine if the interaction of reading speed with the linguistic features has an significant effect on fixation duration. We find the logarithmic token frequency to have a significant interaction with both slow (t = -2.88 and p = 0.004) and moderate readers (t = -2.41 and p = 0.02). Furthermore, we observe that the interaction of the reading speed category and the number of passes a participant makes has a significant effect on fixation duration (p < .001 for both slow and moderate readers).

Focusing on resumption lag times, we observe that, on average, participants took 2.81s (SD = 1.54, range 0.95–8.29) to resume reading after they were interrupted. Among them, five participants have significantly higher resumption lag times compared to the mean resumption lag time (with all having p < .05).

5.1. Discussion

In our analyses on InteRead, we follow established methodologies to validate and examine its characteristics. The analysis of the effects of interruptions on gaze behavior during reading demonstrates a substantial increase in reading times, fixation count and saccade length following an interruption. These findings are in agreement with prior studies, indicating that readers utilize text to restore previously read information upon resumption, as indicated by (Cane et al., 2012). Furthermore, we do not observe a substantial variation in fixation duration and regression frequency between pre- and post-interruption phases. These results, in contrast to Cane et al. (2012), underscore the variability in how interruptions may affect reading behavior. This disparity could be attributed to differences in experimental design, interruption characteristics, or participant-specific factors. Further research is required to delve into the reasons for this discrepancy and enhance our understanding of how interruptions influence gaze behavior during reading.

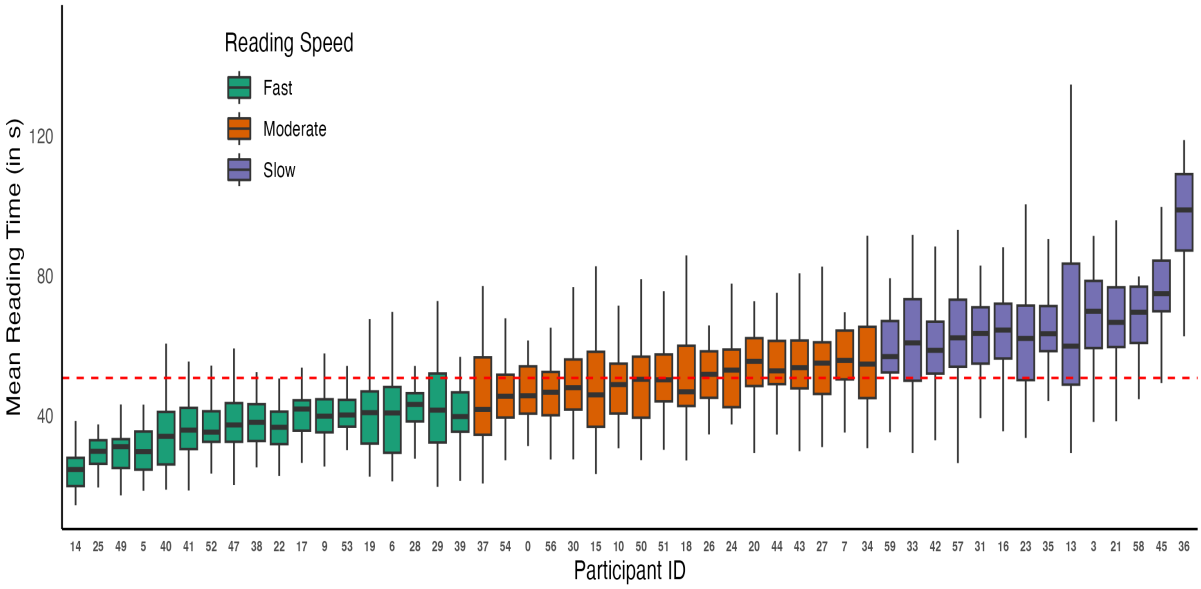

Figure 4: Boxplot showing individual differences between participant (x-axis) and the mean reading time for a page (y-axis). The reading times are sorted in ascending order and the dashed horizontal line denotes the mean reading time for all participants. The participants are grouped into 3 categories based on their reading speed: Fast, Moderate and Slow.

Furthermore, our examination of word length and frequency's effects on gaze features related to early cognitive processes aligns with previous findings (Kliegl et al., 2004) and demonstrate that the influence of word length becomes apparent only after the entire word has been read. Conversely, the effect of word frequency, already significant during the first pass fixation, diminishes once the entire word is read, leading to variations in gaze duration. Notably, gaze duration tends to increase when readers encounter longer words, while they tend to skip highly frequent words (Kliegl et al., 2004). In addition, our analyses show that word skipping probability is significantly influenced by part-of-speech tags, dependency relations, and the semantic value of a word. This highlights how lexical processing of token's function in a sentence plays a substantial role in determining whether it is likely to be skipped (Morris, 1994; Brysbaert and Vitu, 1998).

Lastly, our analysis identify substantial variations in reading speed across the corpus, highlighting individual differences among fast, moderate, and slow readers. Similar variations are also found in Rayner et al. (2010). Individual differences are also observed in resumption lag times, shedding light on the importance of further investigating individual differences in task resumption performance and strategy (Werner et al., 2011).

6. Outlook

Our interdisciplinary analyses showcase that InteRead can be utilized across various research domains, such as educational sciences, psycholinguistics and NLP. In educational sciences, it can facilitate the investigation of the challenges learners face when their attention is disrupted during reading tasks. Future studies could use InteRead to improve the development and evaluation of adaptive AALT that dynamically respond to individual differences in how learners cope with interruptions during reading. For example, by further analyzing differences in gaze behaviors upon resumption, systems could be designed to provide personalized interventions, such as content reminders or visual cues, depending on a student's specific needs and reading habits, as suggested by Mariakakis et al. (2015); Hutt et al. (2021).

In reading research and psycholinguistics, researchers can investigate how external interruptions impact the reading processes and whether resumption strategies employed by readers vary depending on in individual differences. This could lead to a deeper understanding of the interplay between attention and reading comprehension. As there is currently no consensus on whether interruptions have a negative impact on reading comprehension (Chevet et al., 2022), this dataset could be used to further study interrupted reading. Moreover, it allows for further analyzing differences in linguistic phenomena in normal as opposed to disrupted reading.

Within NLP, InteRead could benefit the development of NLP-based models capable of predicting reading performance and attentional states in real-time. For instance, NLP models might be en-

hanced to predict the optimal timing and design for delivering interventions during reading, ensuring they have the greatest impact on maintaining readers' comprehension. In particular, these models might benefit from an input that take into account the disruptions encountered in real-world settings.

7. Acknowledgments

This work was supported by the Federal Ministry of Science, Research, and the Arts Baden-Württemberg as part of the Research Seed Capital funding scheme (grant number Az. 33-7533.-30-10/75/3). We acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation under the DFG reference number UP 31/1) for the Stuttgart Research Focus Interchange Forum for Reflecting on Intelligent Systems (IRIS). E. Sood was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016.
The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Francesca Zermiani. Lastly, we would like to thank the anonymous reviewers for their valuable feedback.
Credit authorship contribution statement: F. Zermiani: Conceptualization, Investigation, Methodology, Writing- Original draft. P. Dhar: Formal Analysis, Validation, Visualization, Writing- Original draft. E. Sood: Supervision, Conceptualization, Methodology, Writing- Reviewing & Editing. F. Kögel: Data Curation, Visualization, Writing- Reviewing & Editing. A. Bulling: Supervision, Writing- Reviewing & Editing. M. Wirzberger: Supervision, Resources, Funding Acquisition, Writing- Reviewing & Editing.

8. Ethical and Broader Impact Statement

Firstly, our corpus predominantly comprises participants from specific demographic backgrounds, gender and ethnicity, potentially limiting the representativity of our data. Future gaze data collection efforts should prioritize inclusivity, seeking to encompass individuals from underrepresented ethnicities and diverse backgrounds.
Moreover, the generalizability of our data should be considered in the context of controlled reading tasks conducted within a laboratory setting, which did not allow for a continuous observation of the participants. However, despite introducing a certain degree of drift in the data due to potential head movements while typing, the controlled interruptions enabled the researchers to reproduce the intricacies and disruptions that student experience in a classroom environment. Reading behavior can vary significantly based on external factors such as text genre, language, and reading purpose, which should be taken into account when making use of InteRead.
A limitation of our analyses stems from the use of particular statistical tests such as ANOVA and linear regression. Both of these tests assume that the variables being evaluated are normally distributed, which was not the case with the fixation features. We nonetheless reported the significance scores using ANOVA and linear regression so as to replicate prior studies that involved fixations.
Lastly, despite of the efforts to ensure annotation consistency, manual annotations of resumption lag times may present inherent subjectivity, potentially introducing slight variations in our dataset. In summary, these limitations underscore the necessity of addressing the inclusion of underrepresented ethnicities, refining recording setups, and recognizing the context-specific nature of eye tracking data for responsible interpretation and application in diverse research contexts.

9. Bibliographical References

Lidia Altamura, Ladislao Salmerón, and Yvonne Kammerer. 2022. Instant messaging multitasking while reading: A pilot eye-tracking study. In *2022 Symposium on Eye Tracking Research and Applications*, ETRA '22, pages 1–6. Association for Computing Machinery.

Erik Altmann and J.Gregory Trafton. 2002. Memory for goals: An activation-based model. *Cognitive Science*, 26:39–83.

Erik M. Altmann and J. Gregory Trafton. 2007. Timecourse of recovery from task interruption: Data and a model. *Psychonomic Bulletin & Review*, 14:1079–1084.

Julia Armendariz, Carla Tamayo, Justin Slade, Ilana Belitskaya-Lévy, Caroline Gray, and Nazima Allaudeen. 2021. Interruptions to attending physician rounds and their effect on resident education. *Journal of Graduate Medical Education*, 13:266–275.

Handan Atun. 2020. Intelligent tutoring systems (ITS) to improve reading comprehension: a systematic review. *Journal of Teacher Education and Lifelong Learning*, 2:77–89.

Hao Bai, Winston E. Jones, Jarrod Moss, and Stephanie M. Doane. 2014. Relating individual differences in cognitive ability and strategy consistency to interruption recovery during multitasking. *Learning and Individual Differences*, 35:22–33.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14:1–16.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Marc Brysbaert and Françoise Vitu. 1998. Word skipping: Implications for theories of eye movement control in reading. In *Eye guidance in reading and scene perception*, pages 125–147. Elsevier.

James E. Cane, Fabrice Cauchard, and Ulrich W. Weger. 2012. The time-course of recovery from interruption during reading: Eye movement evidence for the role of interruption lag and spatial memory. *Quarterly Journal of Experimental Psychology*, 65:1397–1413.

Fabrice Cauchard, James Cane, and Ulrich Weger. 2012. Influence of background speech and music in interrupted reading: An eye-tracking study. *Applied Cognitive Psychology*, 26:381 – 390.

Guillaume Chevet, Thierry Baccino, Lucas Marlot, Annie Vinter, and Véronique Drai-Zerbib. 2022. Effects of interruption on eye movements and comprehension during reading on digital devices. *Learning and Instruction*, 80:101565.

Cristina Conati, Vincent Aleven, and Antonija Mitrovic. 2013. Eye-tracking for student modelling in intelligent tutoring systems. *Design Recommendations for Intelligent Tutoring Systems*, 1:227–236.

Edwin S. Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*, 46:913–921.

Asim H. Dar, Adina S. Wagner, and Michael Hanke. 2020. Remodnav: Robust eye-movement classification for dynamic stimulation. *bioRxiv*, 53:399–414.

Sidney D'Mello, Kristopher Kopp, Robert Earl Bixler, and Nigel Bosch. 2016. Attending to attention: Detecting and combating mind wandering during computerized reading. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 1661–1669. Association for Computing Machinery.

Sidney D'Mello. 2019. Gaze-based attention-aware cyberlearning technologies: Learning in the age of emerging technologies. In Thomas D. Parsons, Lin Lin, and Deborah Cockerham, editors, *Mind, Brain and Technology*, pages 87–105.

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70:377–398.

Amy M. Elleman and Eric L. Oslund. 2019. Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences*, 6:3–11.

Ralf Engbert, Andre Longtin, and Reinhold Kliegl. 2002. A dynamic model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42:621–636.

Cyrus K. Foroughi, Daniela Barragán, and Deborah A. Boehm-Davis. 2016. Interrupted reading and working memory capacity. *Journal of Applied Research in Memory and Cognition*, 5:395–400.

Jeffrey L. Foster, Zach Shipstead, Tyler L. Harrison, Kenny L. Hicks, Thomas S. Redick, and Randall W. Engle. 2015. Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, 43:226–236.

Matthias Gamer, Jim Lemon, and Ian Singh. 2010. *irr: Various Coefficients of Interrater Reliability and Agreement*.

Kevin A. Gluck, John R. Anderson, and Scott A. Douglass. 2000. Broader bandwidth in student modeling: What if ITS were "eye"TS? In *Intelligent Tutoring Systems*, ITS 2000, pages 504–513.

Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. 2022. Patterns of text readability in human and predicted eye movements. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 1–15. Association for Computational Linguistics.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, LREC, pages 138–146. European Language Resources Association.

Stephen Hutt, Kristina Krasich, James R. Brockmole, and Sidney D'Mello. 2021. Breaking

out of the lab: Mitigating mind wandering with gaze-based attention-aware technology in classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–14. Association for Computing Machinery.

Stephen Hutt, Caitlin Mills, Shelby White, Patrick J. Donnelly, and Sidney D'Mello. 2016. The Eyes Have It: Gaze-based detection of mind wandering during learning with an intelligent tutoring system. In *Proceedings of the 9th International Conference on Educational Data Mining*, EDM 2016, pages 86–93. International Educational Data Mining Society.

Albrecht W. Inhoff and Ulrich W. Weger. 2005. Memory for word location during reading: Eye movements to previously read words are spatially selective but not precise. *Memory & Cognition*, 33:447–461.

Matthew E. Jacovina and Danielle S. McNamara. 2017. Intelligent tutoring systems for literacy: Existing technologies and continuing challenges. In R. Atkinson, editor, *Intelligent Tutoring Systems: Structure, Applications and Challenges*, pages 153–174. Nova Science Publishers Inc, Hauppauge, NY.

Jaemin Jo, Bohyoung Kim, and Jinwook Seo. 2015. EyeBookmark: Assisting recovery from interruption during reading. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*, CHI '15, pages 2963–2966. Association for Computing Machinery.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87:329.

Byeongwoo Kang. 2021. How the COVID-19 pandemic is reshaping the education service. *The Future of Service Post-COVID-19 Pandemic, Volume 1: Rapid Adoption of Digital Service Technology*, pages 15–36.

Izabela Krejtz, Agnieszka Szarkowska, and Maria Łogińska. 2016. Reading function and content words in subtitled videos. *The Journal of Deaf Studies and Deaf Education*, 21:222–232.

Moniek Kuijpers, Frank Hakemulder, Ed Tan, and Miruna Doicaru. 2014. Exploring absorbing reading experiences: Developing and validating a self-report scale to measure story world absorption. *Scientific Study of Literature*, 4:89–122.

Kai Kunze, Andreas Bulling, Yuzuko Utsumi, Shiga Yuki, and Koichi Kise. 2013. I know what you are reading – Recognition of document types using mobile eye tracking. In *Proc. IEEE International Symposium on Wearable Computers (ISWC)*, pages 113–116.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Diane Litman. 2016. Natural language processing for enhancing teaching and learning. In *Proceedings of the 2016 AAAI Conference on Artificial Intelligence*, AAAI'16, pages 4170–4176. AAAI Press.

Alexander Mariakakis, Mayank Goel, Md Tanvir Islam Aumi, Shwetak N. Patel, and Jacob O. Wobbrock. 2015. SwitchBack: Using focus and saccade tracking to guide users' attention for mobile task resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2953–2962. Association for Computing Machinery.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharya. 2020. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4907–4913. International Joint Conferences on Artificial Intelligence Organization.

Hilkka L. Meys and Penelope M. Sanderson. 2013. The effect of individual differences on how people handle interruptions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 868–872.

Diane Mézière, Lili Yu, Erik D. Reichle, Titus von der Malsburg, and Genevieve M McArthur. 2023. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58:425–449.

Caitlin Mills, Julie Gregg, Robert Bixler, and Sidney K. D'Mello. 2021. Eye-Mind reader: an intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering. *Human–Computer Interaction*, 36:306–332.

Robin K Morris. 1994. Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:92–103.

Moshe Naveh-Benjamin, Jonathan Guez, and Shai Sorek. 2007. The effects of divided attention on encoding processes in memory: Mapping the locus of interference. *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Expérimentale*, 61:1–12.

Harold Pashler, Sean H. K. Kang, and Renita Y. Ip. 2013. Does multitasking impair studying? Depends on timing. *Applied Cognitive Psychology*, 27:593–599.

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51:195–203.

Leslie Gross Portney, Mary P Watkins, et al. 2009. *Foundations of clinical research: applications to practice*, volume 892. Pearson/Prentice Hall Upper Saddle River, NJ.

Cassandra Potier Watkins, Julien Caporal, Clément Merville, Sid Kouider, and Stanislas Dehaene. 2020. Accelerating reading acquisition and boosting comprehension with a cognitive science-based tablet training. *Journal of Computers in Education*, 7:183–212.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422.

Keith Rayner. 2009. Eye movements in reading: Models and data. *Journal of Eye Movement Research*, 2:1–10.

Keith Rayner, Kathryn Chace, Timothy Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10:241–255.

Keith Rayner and George W. McConkie. 1976. What guides a reader's eye movements? *Vision Research*, 16:829–837.

Keith Rayner, Sara C Sereno, Robin K Morris, A Rene Schmauder, and Charles Clifton Jr. 1989. Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4:SI21–SI49.

Keith Rayner, Timothy Slattery, and Nathalie Bélanger. 2010. Eye movements, the perceptual span, and reading speed. *Psychonomic Bulletin & Review*, 17:834–839.

Thomas S. Redick, James M. Broadway, Matt E. Meier, Princy S. Kuriakose, Nash Unsworth, Michael J. Kane, and Randall W. Engle. 2012. Measuring working memory capacity with automated complex span tasks. 28:164–171.

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological Review*, 105:125.

Erik D Reichle, Andrew E Reineberg, and Jonathan W Schooler. 2010. Eye movements during mindless reading. *Psychological Science*, 21:1300–1310.

Rufat Rzayev, Paweł W Woźniak, Tilman Dingler, and Niels Henze. 2018. Reading on smart glasses: The effect of text position, presentation type and walking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–9.

Hildur Schilling, Keith Rayner, and James Chumbley. 1998. Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, 26:1270–1281.

Priti Shah, Akira Miyake, and Akira Miyake. 1996. The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology. General*, 125:4–27.

John L. Sibert, Mehmet Gokturk, and Robert A. Lavine. 2000. The reading assistant: Eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology*, UIST '00, page 101–107. Association for Computing Machinery.

Reid Smith, Pamela Snow, Tanya Serry, and Lorraine Hammond. 2021. The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, 42:214–240.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. In *Proc. ACL SIGNLL Conference on Computational Natural Language Learning*, CoNLL, pages 12–25. Association for Computational Linguistics.

Namrata Srivastava, Rajiv Jain, Jennifer Healey, Zoya Bylinskii, and Tilman Dingler. 2021. Mitigating the effects of reading interruptions by providing reviews and previews. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–6. Association for Computing Machinery.

J. Gregory Trafton, Erik M. Altmann, Derek P. Brock, and Farilee E. Mintz. 2003. Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, 58:583–603.

Nash Unsworth, Thomas S. Redick, Richard P. Heitz, James M. Broadway, and Randall W. Engle. 2009. Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, 17:635–654.

Nicole E. Werner, David M. Cades, Deborah A. Boehm-Davis, Jessica Chang, Hibah Khan, and Gia Thi. 2011. What makes us resilient to interruptions? Understanding the role of individual differences in resumption. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 55, pages 296–300.

Kausalai Kay Wijekumar, Bonnie JF Meyer, and Puiwa Lei. 2012. Large-scale randomized controlled trial with 4th graders using intelligent tutoring of the structure strategy to improve nonfiction reading comprehension. *Educational Technology Research and Development*, 60:987–1013.

Maria Wirzberger and Günter Daniel Rey. 2018. Attention please! Enhanced attention control abilities compensate for instructional impairments in multimedia learning. *Journal of Computers in Education*, 5:243–257.

Maria Wirzberger and Nele Russwinkel. 2015. Modeling interruption and resumption in a smartphone task: An ACT-R approach. *i-com*, 14(2):147–154.

10. Language Resource References

Brysbaert, Marc and New, Boris. 2009. *Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English*. [link].

Brysbaert, Marc and Warriner, Amy and Kuperman, Victor. 2013. *Concreteness ratings for 40 thousand generally known English word lemmas*. [link].

Cop, Uschi and Dirix, Nicolas and Drieghe, Denis and Duyck, Wouter. 2017. *Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading*. [link].

Frank, Stefan and Monsalve, Irene and Thompson, Robin and Vigliocco, Gabriella. 2013. *Reading time data for evaluating broad-coverage models of English sentence processing*. [link].

Nora Hollenstein and Maria Barrett and Marina Bjornsd'ottir. 2022. *The Copenhagen Corpus of Eye Tracking Recordings from Natural Reading of Danish Texts*. European Language Resources Association, LREC 2022. [link].

Nora Hollenstein and Marius Troendle and Ce Zhang and Nicolas Langer. 2020. *ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation*. European Language Resources Association, LREC. [link].

Honnibal, Matthew and Montani, Ines and Van Landeghem, Sofie and Boyd, Adriane. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. [link].

Samar Husain and Shravan Vasishth and Narayanan Srinivasan. 2015. *Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus*. [link].

Jäger, Lena and Kern, Thomas and Haller, Patrick. 2021. *Potsdam Textbook Corpus (PoTeC): Eye tracking data from experts and non-experts reading scientific texts*. [link].

Kliegl, Reinhold and Grabner, Ellen and Rolfs, Martin and Engbert, Ralf. 2004. *Length, frequency, and predictability effects of words on eye movements in reading*. Routledge. [link].

Kuperman, Victor and Siegelman, Noam and Schroeder, Sascha and Acartürk, Cengiz and Alexeeva, Svetlana and Amenta, Simona and Bertram, Raymond and Bonandrini, Rolando and Brysbaert, Marc and Chernova, Daria and others. 2023. *Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus*. [link].

Luke, Steven and Christianson, Kiel. 2018. *The Provo Corpus: A large eye-tracking corpus with predictability norms*. [link].

Mishra, Abhijit and Kanojia, Diptesh and Bhattacharyya, Pushpak. 2016. *Predicting Readers' Sarcasm Understandability by Modeling Gaze Behavior*. AAAI Press, AAAI '16'. [link].

Pan, Jinger and Yan, Ming and Richter, Eike M and Shu, Hua and Kliegl, Reinhold. 2022. *The Beijing Sentence Corpus: A Chinese sentence corpus with eye movement data and predictability norms*. [link].

Molood Sadat Safavi and Samar Husain and Shravan Vasishth. 2016. *Dependency Resolution Difficulty Increases with Distance in Persian Separable Complex Predicates: Evidence for Expectation and Memory-Based Accounts*. [link].

Siegelman, Noam and Schroeder, Sascha and Acartürk, Cengiz and Ahn, Hee-Don and Alexeeva, Svetlana and Amenta, Simona and Bertram, Raymond and Bonandrini, Rolando and Brysbaert, Marc and Chernova, Daria and others. 2022. *Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO)*. [link].

Sood, Ekta and Kögel, Fabian and Strohm, Florian and Dhar, Prajit and Bulling, Andreas. 2021. *VQA-MHUG: A gaze dataset to study multimodal neural attention in VQA*. Association for Computational Linguistics. [link].

Yi, Kun and Guo, Yu and Jiang, Weifeng and Wang, Zhi and Sun, Lifeng. 2020. *A Dataset for Exploring Gaze Behaviors in Text Summarization*. Association for Computing Machinery, MMSys '20. [link].

Appendices

Appendix A. Pre-Test Questionnaire

The pre-test included a demographic questionnaire, a prior knowledge questionnaire, and the shortened version of a symmetry span task (SSPAN).

Demographic Questionnaire For multiple-choice questions, each option was labeled with a number. All labels are reported within parentheses here and were not presented to the participants.

1. Please, type your participant ID code:
2. What is your age? Please type your answer:
3. What is your gender? (0) Female, (1) Male, (2) Other, (3) Prefer not to answer
4. Are you a native English speaker? (0) No, (1) Yes UK English, (2) Yes US English
5. Please indicate which is your dominant hand: (0) Both, (1) Right, (2) Left
6. Do you require eye correction? If yes, which one are you wearing now? (0) No, (1) Yes - contact lenses, (2) Yes - glasses
7. What is your educational background or current study subject/program?
8. Please indicate your level of formal education, indicating the degrees which are in progress (i.e., bachelors degree in progress, bachelors degree, master in progress, master, PhD in progress, etc.)
9. Would you consider yourself a speed reader (i.e., someone who generally reads and understands text faster)? (0) No, (1) Yes
10. How often do you read for enjoyment? (0)Never, (1) 1-2 times a week, (2) 2-3 times a week, (3) 4-5 times a week, (4) Everyday
11. What medium do you choose to read from? (0) Print book, (1) E-book, (2) Computer, (3) Smartphone, (4) Other

Prior Knowledge Questionnaire We designed this questionnaire to assess participants' prior knowledge on Sherlock Holmes stories as well as familiarity with crime fiction on a broader level. To prevent neutral responses, a 6-point Likert scale was selected for all Likert scale questions administered during data collection. The questionnaire can be grouped in three subsets of questions: *detailed knowledge* (1-4) with multiple-choice questions concerning details of Sherlock Holmes stories, all including the option 'I do not know' to refrain participants from guessing the correct answer; *target-domain knowledge* (5-8) covering yes/no agreement statements about prior exposure to crime fiction, TV adaptation of Sherlock Holmes, the selected story used in the study, or other Sherlock Holmes stories; *general knowledge* (9-12) including 6-point Likert scale agreement statements about participants' interest in the detective or crime genre. The scores assigned are enclosed in parentheses and their cumulative sum constitutes the prior knowledge total score. We evaluated the internal consistency of our prior knowledge questionnaire and obtained an overall Cronbach's Alpha of α = .83, with detailed knowledge (α = .68), target-domain knowledge (α = .45), and general knowledge (α = .90).

1. Who is the author of Sherlock Holmes? (0) Virginia Wolf, (0) Edgard Allan Poe, (1) Arthur C. Doyle, (0) Agatha Christie, (0) I do not know
2. Which of the following statement is true about Sherlock Holmes stories? (0) The stories usually begin at home where Holmes and Watson live, (0) The crime is described in details by the client, (0) Holmes and Watson investigate the location of the crime and collect clues, (1) All of the above are true, (0) I do not know
3. What is the name of Sherlock Holmes' landlady? (0) Mrs Hudson, (0) Mrs Hudley, (0) Mrs Whitfield, (0) Mrs Haden, (0) I do not know
4. Where do Sherlock Holmes and Watson live? (0) High Street, (1) Baker Street, (0) Regency Street, (0) Oxford Street, (0) I do not know

5. I have recently read a detective/crime fiction story (for example, works by Agatha Christie / Arthur C. Doyle / Stieg Larsson / Jo Nesbø / Camilla Läckberg, or other crime-fiction writers). (1) Yes, (0) No

6. I have recently seen a film or a TV adaptation of Sherlock Holmes (for example, the BBC series Sherlock). (1) Yes, (0) No

7. I have already read the Sherlock Holmes story "The Adventure of the Speckled Band' before. (1) Yes, (0) No

8. I have read other Sherlock Holmes stories before. (1) Yes, (0) No

9. I read detective/crime fiction (for example, works by Agatha Christie / Arthur C. Doyle / Stieg Larsson / Jo Nesbø / Camilla Läckberg, or other crime-fiction) with great interest. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree

10. I usually recommend reading detective/crime fiction books. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree

11. When deciding for a new book to read, I often pick a detective/crime fiction story. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree

12. I enjoy reading detective/crime fiction (for example, works by Agatha Christie / Arthur C. Doyle / Stieg Larsson / Jo Nesbø / Camilla Läckberg, or other crime-fiction writers). (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree

Symmetry Span Task The SSPAN was first developed by (Shah et al., 1996) to target spatial working memory capacity and spatial visualization skills. We used the short version of the SSPAN (Foster et al., 2015). The main task requires participants to remember the positions of colored squares in a 4x4 grid in the correct sequence, after each square appears for 650 ms. The distractor task instead consists of providing symmetry judgments on a series of grids containing colored squares, determining whether they are vertically symmetrical or not. The number of squares can vary from two to five per trial according to Foster et al. (2015). We administered three blocks of trials in random order, to prevent participants from developing an understanding of how many red squares they will need to recall. The participants had three rounds of practice to familiarize themselves with each task (Redick et al., 2012): (i) main task, (ii) distractor task, (iii) both main and distractor tasks combined. To assign scores, we followed an established procedure for cognitive span tasks and considered the partial recall score, which - in the case of the SSPAN - consists of the sum of red squares recalled in the correct location, independent of whether all the squares in each trial were correctly recalled in the right order (Redick et al., 2012; Foster et al., 2015). The partial recall score ranges from 0 to 42 points per individual (Unsworth et al., 2009), and it has been shown to provide more robust and internally consistent scores (Redick et al., 2012). The dataset includes: the partial recall score, the total number of correct symmetry judgments, and the total accuracy (i.e., the proportion of the correct symmetry judgments).

Appendix B. Post-Test Questionnaire

The post-test included a short reading comprehension questionnaire and a set of 6-point Likert scale statements targeting the participants' reading experience and the strategy they used to re-focus on reading after being interrupted.

Reading Comprehension Questionnaire The questionnaire consisted of four multiple-choice inferential questions to examine participants' engagement with the text. The scores assigned are enclosed in parentheses and their cumulative sum constitutes the reading comprehension total score.

1. Why did Miss Helen Stoner decide to ask Sherlock Holmes' advice? (1) Because she suddenly heard that same low whistle her sister heard before dying, (0) Because she was cruelly treated by her stepfather, (0) Because she had to move into the room where her sister died, (0) Because her friend suggested that she should ask him for advice

2. What do you think the 'speckled band' is? (0) It is just a made-up phrase Julia Stoner said out of shock, (0) It is a phrase Julia Stoner said to refer to the gipsies in the plantation, (0) It is a phrase Julia Stoner said to refer to the cheetah's fur, (1) It is a phrase Julia Stoner said to refer to a venomous snake

3. What is the purpose of Dr. Roylott's visit? (0) He wants to intimidate Sherlock Holmes and Dr. Watson, (0) He wants to threaten Dr. Watson not to interfere with his business, (1) He wants to threaten Sherlock Holmes not to stick his nose into his affairs, (0) He wants to beat Sherlock Holmes and Dr. Watson

4. Who do you think the murderer is? (0) One of the gipsies from the plantation, (0) Julia Stoner's fiancé, (1) Dr. Roylott, (0) Miss Helen Stoner's fiancé

Reading Experience Questionnaire This questionnaire assessed participants' interest in the story and their reading experience (1-6). Among these, four questions (2-5) were selected from the story world absorption scale (Kuijpers et al., 2014) and adapted to a 6-point Likert scale. Finally, building on previous findings (Wirzberger and Russwinkel, 2015), participants were asked two questions (7-8) concerning the strategy they used to refocus on the story after each interruption.

1. I was annoyed by the interruptions. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree
2. When I was reading the story I was focused on what happened in the story. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree
3. When I was reading the story I had an image of the main characters in mind. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree
4. Something in the story stuck with me after I finished reading it. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree
5. I read the story with great interest. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree
6. I felt completely occupied by the story despite of the interruptions. (0) Strongly disagree, (1) Disagree, (2) Slightly disagree, (3) Slightly agree, (4) Agree, (5) Strongly agree
7. After each interruption, what did you do to refocus on the story? (0) I searched for the exact point in the text I was reading before the interruption (searching), (1) I tried to remember what content I was reading before the interruption (remembering), (2) Both
8. Did you re-focus on the story in other ways? Please type your answer

Appendix C. Complete Statistical Models

We report here the complete statistical models used in the following sections of our paper. For all linear mixed effects regression analyses, we employ the lme4 package (version 1.1-35.1) (Bates et al., 2015). To extract the p-values, which are not included in the lme4 output, we use the lmerTest package (version 3.1-3) (Kuznetsova et al., 2017).

Word Length and Word Frequency Effect on Gaze Behavior The full models used here are:

```
lmer(first_pass~word_len+log_freq+(1|page_id)
+ (1|line_num) + (1|participant_id),data=df)

lmer(gaze_dur~word_len+log_freq+(1|page_id)
+ (1|line_num) + (1|participant_id),data=df)
```

Individual Differences in Reading and Resumption Time The full models used here are:

```
lmer(fix_dur~reading_cat*log_freq+(1|page_id)
+ (1|line_num) + (1|participant_id),data=df)

lmer(fix_dur~pas_num+log_freq+(1|page_id)
+ (1|line_num) + (1|participant_id),data=df)
```