Eye Movements in Information-Seeking Reading

Omer Shubi[1] **(shubi@campus.technion.ac.il)**
Yevgeni Berzak[1,2]**(berzak@technion.ac.il)**
[1] Faculty of Data and Decision Sciences, Technion - Israel Institute of Technology, Haifa, Israel
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA

Abstract

In this work, we use question answering as a general framework for studying how eye movements in reading reflect the reader's goals, how they are pursued, and the extent to which they are achieved. We leverage fine-grained annotations of task-critical textual information to perform a detailed comparison of eye movements in information-seeking and ordinary reading regimes. We further examine how eye movements during information seeking relate to question answering behavior. We find that reading times, saccade patterns and sensitivity to the linguistic properties of the text are all strongly and systematically conditioned on the reading task, and further interact with question answering behavior. The observed reading patterns are consistent with a rational account of cognitive resource allocation during task-based reading.

Keywords: eye movements, reading, question answering

Introduction

It has long been established that eye movements in reading contain rich information on how readers comprehend language on a moment-by-moment basis (Just & Carpenter, 1980; Rayner, 1998). However, the exact relations between eye movements in reading and language comprehension are still far from well understood. Furthermore, the large majority of eye movements research examined ordinary reading, in which the reader is not given a specific goal beyond general comprehension of the text. Such research leaves out many daily situations in which readers are guided by concrete goals with respect to the text. These goals often involve seeking specific information of interest, and can influence eye movement patterns in reading. An account of goal driven reading has therefore been acknowledged as being essential for developing general models for eye movements in reading (Radach & Kennedy, 2004).

Here, we take a step in this direction by proposing a general framework for studying how readers' goals influence their reading patterns in naturalistic reading. We operationalize a goal as seeking an answer to a question about the text. Using this framework we follow Hahn & Keller (2023) and Malmaud et al. (2020) and examine two reading regimes, an information-seeking "Hunting" regime in which the question is presented to participants prior to reading the passage, and an ordinary reading "Gathering" regime where participants are presented with the question only after having read the passage. Our study provides a fine-grained characterization of the similarities and differences in reading patterns between these two regimes, examining reading times, regressive saccades, and response to linguistic characteristics of the text as participants progress through the passage.

Our analyses are enabled by OneStopGaze (Malmaud et al., 2020), a broad coverage eye movements in reading dataset with 269 participants reading text passages and answering multiple choice reading comprehension questions about them. A key characteristic of OneStopGaze is the underlying textual annotations. The annotations structure answer choices by degree of reading comprehension and tie them to their textual support (Berzak et al., 2020). Importantly, the annotations mark the *Critical Span*: the part of the text which contains the relevant information for answering the question correctly. We center our analysis on the resulting division between task-critical information, and the information that precedes and follows it. Overall, the textual annotations enable analyses of the relations between eye movements, reading goals, and reading comprehension behavior, which has thus far been limited using existing eye movement datasets.

Our analyses yield the following key conclusions:

1. In information seeking reading regimes readers engage differently with both task relevant and task irrelevant information compared to ordinary reading, exhibiting highly strategic and cognitive resource efficient reading patterns.

2. The extent to which (1) holds diminishes when readers are not successful in solving the reading comprehension task.

These results inform our understanding of task-based reading, as well as the relation between eye movements in reading and reading comprehension.

Related Work

Task-based reading has been studied using two broad approaches which we refer to as *procedural-based* and *content-based*. Both approaches can be thought of as providing the reader with a prompt prior to reading the target text. The first approach consists of tasks that specify a reading *procedure*, and are typically not derived directly from the specific content of the text. As such, they are often general, and widely applicable across texts. At the same time, this approach is limited in the range of tasks, with most studies focusing on a small number of canonical tasks. A key early study within this framework is Just et al. (1982), who compared ordinary reading, skimming, and speed reading with respect to reading

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

speed, fixation durations, and reading comprehension performance. J. Kaakinen & Hyönä (2010) compared ordinary reading to proofreading. Among others, the proofreading regime was characterized by shorter saccades, longer fixations, and larger word length and word frequency effects. Schotter et al. (2014) also compared ordinary reading to proofreading, and similarly found larger frequency effects in proofreading. They further observed larger predictability effects when proofreading involves real but unintended words, the identification of which requires contextual integration. Rayner & Raney (1996) examined differences between ordinary reading and searching through the text for a target word, and found no frequency effects in the latter. Several studies have also examined eye movements during human linguistic annotation, often used for generating training data for Natural Language Processing (NLP) tools, such as annotation of named entities (Tomanek et al., 2010; Tokunaga et al., 2017) and semantic relations (Hollenstein et al., 2021, 2022).

In the second line of research, which we extend, tasks are derived more explicitly from the *content* of the text. In this approach, tasks are constrained in that they have to be specific to a given text, but can vary otherwise, allowing the possibility to incorporate a large number of tasks within and across studies. In this framework, differences in reading patterns were found when readers were asked to take different perspectives on a given text (J. K. Kaakinen et al., 2002), with longer fixation times on perspective-relevant than perspective-irrelevant information. Rothkopf & Billington (1979) examined task-based reading where participants are given a set of learning goals, formulated as reading comprehension questions. They found more and longer fixations on task-relevant than task-irrelevant text, and overall shorter reading times than in ordinary reading. Our findings are generally consistent with these results.

Our work is closest to Hahn & Keller (2023) and Malmaud et al. (2020) who examine how eye movements are conditioned on a single reading comprehension question. Hahn & Keller (2023) collected eye-tracking data for materials from the CNN and Daily Mail corpus which contains questions whose answer is a named entity (Hermann et al., 2015). They demonstrate that reading times on the named entity which is the correct answer to the question are longer if participants are shown the question before reading the passage as compared to ordinary reading. Malmaud et al. (2020) generalize this experimental setup using the OneStopQA corpus (Berzak et al., 2020), which has arbitrary questions whose answer can be inferred from a well-defined span in the text, called the *Critical Span*, as mentioned above. Differently from ordinary reading, when participants are shown the question before reading the passage, mean word reading times were found to be longer within the Critical Span than outside it. Further, consistent with findings by Rothkopf & Billington (1979) and Hahn & Keller (2023), the overall reading time in the Hunting condition was found to be shorter than in the Gathering condition, driven by shorter reading times outside the Critical Span.

We extend the work of Malmaud et al. (2020) in a number of ways. First, while in their work only participants who answered the question correctly were analyzed, here we analyze all the available data. We further perform a finer-grained split than the inside versus outside the Critical Span division done in Malmaud et al. (2020), distinguishing between the region before the Critical Span and after it. Further, we examine a wider variety of eye movement measures, including saccade based information, and analyze how reading patterns evolve as a function of the word position in the passage. Importantly, we include analyses that examine the sensitivity of reading times to linguistic characteristics of the text. This aspect is key for understanding the relation between reading and cognitive state, but has thus far not been examined in the context of task-based reading. Finally, we analyze how task conditioning interacts with reading comprehension. The combination of these factors provides a substantially more detailed picture of the way in which eye movement dynamics in task-based reading unfold over time and how they relate to reading comprehension.

Experimental Setup and Data

We use OneStopQA (Berzak et al., 2020), a multiple-choice reading comprehension dataset that comprises 30 Guardian articles from the OneStopEnglish corpus (Vajjala & Lučić, 2018). Each article is available in three difficulty levels, of which two are used in the eye-tracking experiment described below: Advanced (the original Guardian article) and a simplified Elementary version. Each article has 4-7 paragraphs, each annotated with three multiple-choice reading comprehension questions. In total, the textual data consists of 162 paragraphs, corresponding to 972 unique paragraph–level–question triplets.

Each question has four answers which belong to four answer types ordered by degree of comprehension, and are further tied to manually annotated spans in the passage. **A** is the correct answer, with support in the *Critical Span*. Note that while the Critical Span contains the information essential for answering the question correctly, it does not contain the correct answer in verbatim form. **B** corresponds to a miscomprehension of the Critical Span. **C** refers to another part of the text outside the Critical Span. **D** has no textual support. The average paragraph length in the textual data is 109 words, 32 of which belong to the Critical Span.

Eye movements data were collected by Malmaud et al. (2020) from 269 native English speakers using an EyeLink 1000 Plus eye tracker (SR Research) at a sampling rate of 1000Hz. Each participant read 10 articles comprising 54 paragraphs from both Advanced and Elementary levels and answered a comprehension question about each paragraph in one of two between-subjects question answering conditions: *Hunting* or *Gathering*. Differently from the Gathering condition, in the Hunting condition participants were shown the question (without the answers) prior to reading the paragraph. Except for this difference, the experimental materials and procedure were identical across conditions. See Malmaud et al. (2020) for further details on the eye-tracking experiment. Overall,

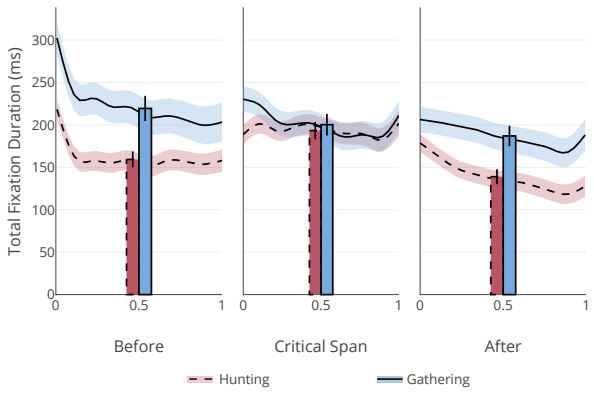

Figure 1: Total Fixation Duration. The x axis represents normalized word position within the corresponding interest area. Curves are GAM fits with random effects for subjects and paragraphs, with 95% confidence intervals. Bars represent per-word interest area averages with 95% confidence intervals.

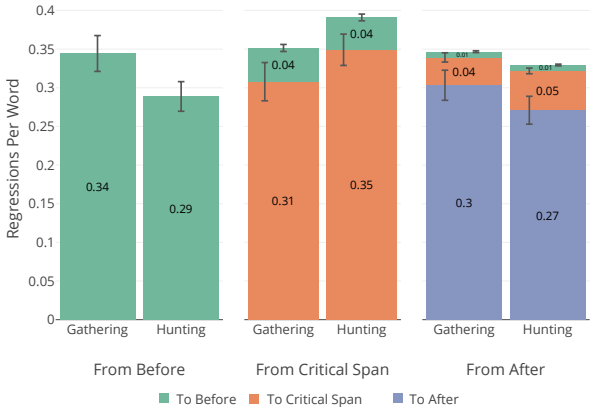

Figure 2: Regression Rate. Number of backward saccades per word with 95% confidence intervals. Each bar is split to regressions from the word that land within the current interest area, and regressions that cross to preceding interest areas.

the eye movements data consists of 7,344 Hunting trials and 7,180 Gathering trials, where a trial is a single response to a reading comprehension question after having read the paragraph.

Information-Seeking versus Ordinary Reading

We characterize differences in eye movement patterns between information-seeking and ordinary reading using two types of information. The first is standard fixation and saccade based measures. The second is the sensitivity of reading times to the linguistic characteristics of the text. Both types of analyses take advantage of the Critical Span annotations to divide each passage into three interest areas: *Before*, *Inside*, and *After* the Critical Span. This division reflects our expectation that task-driven reading behavior may be manifested differently before and after encountering and processing task-critical information. We analyze the reading patterns across the three interest areas and the Hunting and Gathering conditions.

Eye Movement Measures

We first characterize eye movements by examining the following eye movement measures:

- Total Fixation Duration (TF): the sum of all the fixation durations on the word.
- Regression Rate (RR): Number of saccades per word to an earlier part of the paragraph, considering only words that were fixated at least once.

While TF focuses on aggregated *fixations*, RR provides complementary information from the sequence of *saccades*. The Supplemental Material (SM)[1] includes analyses for additional measures, including Skip Rate, first pass First Fixation (FF), Gaze Duration (GD), and number of fixations.

[1] https://osf.io/3gyv9/

We calculate interest area averages of TF times from a mixed effects model[2] applied to each combination of interest area (Before, Inside, After) and reading condition (Hunting, Gathering). To obtain a finer grained view of reading dynamics throughout the paragraph, we further trace these measures as a function of word position within each interest area. We do so using General Additive Models (GAMs), which can fit non-linear relationships between predictors and responses.[3] To account for differences in interest area lengths across items, we normalize word positions to be between 0 (beginning of interest area) and 1 (end of interest area) within each item.

Figure 1 presents interest area averages along with GAM fits for the relation between TF and word position.[4] In the Gathering condition the mean TF times are longer in the Before area than in the Critical Span ($19ms, p < 10^{-11}$), and longer in the Critical Span than in the After area ($14ms, p < 10^{-8}$). In the Hunting condition TF times in the Before area are similarly longer than in the After area ($20ms, p < 10^{-11}$) but crucially are *shorter* than in the Critical Span ($35ms, p < 10^{-19}$). TF times in the Hunting condition are shorter in the Before area ($60ms, p < 10^{-11}$) as well as in the After area ($48ms, p < 10^{-10}$) compared to the corresponding areas in the Gathering condition, but are similar across the conditions inside the Critical Span ($p = 0.361$).

[2] $TF \sim 1 + (1|subj) + (1|parag)$

[3] GAM curves are fitted using the `mgcv` (1.8-41) `bam` function with cubic regression splines (Wood, 2004; Wood et al., 2015).

[4] $TF \sim s(norm\_position, bs = "cr") + s(subj, bs = "re") + s(subj, norm\_position, bs = "re") + s(parag, bs = "re") + s(parag, norm\_position, bs = "re")$, where *norm_position* is the word's position in the interest area, normalized to a 0-1 range. In the absence of visual evidence for non-linearities past the first 15% of the passage, statistical tests are performed using a linear model: $TF \sim norm\_position + (norm\_position|subj) + (norm\_position|parag)$ for each interest area and condition separately. Differences in slopes are tested using the interaction term $norm\_position : condition$ in the following model applied to each interest area $TF \sim norm\_position * condition + (norm\_position * condition|subj) + (norm\_position * condition|parag)$.

Importantly, when examining the progression of TF times as a function of word position, we also observe different patterns of slopes across the reading conditions. In the Gathering condition, TF times decrease within each of the three interest areas ($p < 10^{-10}$) at roughly the same rate across areas. The Hunting slopes exhibit a different behavior. In the Before area, TF times in the Hunting condition decrease slower than in the Gathering condition ($p < 10^{-6}$). In fact, past the initial beginning-of-passage rapid decrease in both conditions (first 15% of the Before area) Hunting TF times remain constant ($p = 0.536$), while Gathering TF times continue to decrease ($p < 10^{-7}$). Inside the Critical Span, Hunting TF times are also constant ($p = 0.053$), with a significant interaction across the conditions ($p < 10^{-5}$). Finally, in the After area, TF times decrease faster than in the Gathering condition ($p = 0.03$). In short, compared to Gathering TF times which decrease at a similar rate across all three sections, Hunting TF times are constant before and within the Critical Span, and decrease faster after the Critical Span.

The Hunting TF patterns suggest an efficient allocation of reading times across interest areas, and deployment of information seeking strategy which consist of three stages around task-critical information: *seek*, *solve*, and *wrap-up*. During the initial *seek* stage readers look for task relevant information while engaging in a skimming like behavior marked by short and constant reading times across word positions. In the *solve* stage readers identify and process task-critical information. In the last stage, the *wrap-up*, reading times are again shorter than in ordinary reading and decrease at a faster rate.

In SM Figure 1 we present a word Skip Rate analysis, which yields a pattern consistent with the TF results. Similar results are also obtained when examining the number of fixations per word in SM Figure 2. We note that this pattern is not fully apparent with early fixation measures which do not aggregate all the fixation times on a word. The SM presents the above analysis for the first pass measures FF (SM Figure 3) and GD (SM Figure 4). In both cases we observe moderately shorter reading times in the Hunting condition, including the Critical Span, without by-area differences within either of the two reading conditions. This suggests that re-reading, especially of words in the Critical Span, is crucial in task-based reading. This conclusion is further supported when examining second and higher pass reading times (SM Figure 5). Not only do they exhibit similar slope patterns to Figure 1, but differently from the early measures have comparable reading times in the Critical Span across the Hunting and Gathering conditions.

To examine this further, we turn to information from *saccades*, focusing on regressions. Figure 2 presents the number of regressions per word within each interest area, broken down by regression landing locations falling into each interest area. The lower Hunting TF times and higher skip rates in the Before area are accompanied by a lower Hunting than Gathering regression rate in this area ($p < 10^{-3}$). However, despite comparable TF times in the Critical Span and lower TF times in the After area in the Hunting condition, we see a higher Hunting regression rate in the Critical Span ($p = 0.02$), and a similar regression rate in the After area ($p = 0.26$).

Furthermore, the breakdown of regression landing locations reveals that the driver of the increased Hunting regression rate in the Critical Span is a higher rate of regressions that land within the Critical Span ($p = 0.009$). Accordingly, compared to the Gathering condition, regressions that are initiated within the Critical Span in the Hunting condition are more likely to land within the Critical Span than to cross to the Before area ($p < 10^{-3}$). At the same time, regressions from the After area are more likely to land in the Critical Span than to remain in the After area as compared to the Gathering condition ($p < 10^{-13}$). Thus, the After area is marked not only by fast decreasing reading times, but also by a stronger tendency to return to task-critical information.

The overall pattern of regressions within and across interest areas suggests that task-driven reading leads to an increased fraction of regressions landing on task-critical information. Additionally, in SM Figure 6 we provide the same analysis for forward saccades, where we see an opposite pattern around the Critical Span: a larger fraction of Hunting forward saccades from the Before area landing in the Critical Span, and a smaller fraction of forward saccades from the Critical Span crossing over to the After area. These patterns are consistent with the seek–solve–wrap-up strategy proposed in the TF times analysis, and provide a more detailed picture on how it arises from the sequence of saccades, with the Critical Span attracting saccades at a higher rate from both the Before and After areas, and retaining saccades at a higher rate than in the Gathering condition.

Response to Linguistic Properties of the Text

It has been widely established that word reading times during ordinary reading are robustly predicted by linguistic properties of words, in particular by word frequency, predictability, and word length (Kliegl et al. (2004); Rayner et al. (2004, 2011) among others). Further, the magnitude of word property effects on reading times has been linked to the reader's cognitive state (Reichle et al., 2010) and language proficiency (Berzak et al., 2022). However, thus far word property effects have not been examined in the context of task-based reading.

Here, we examine the extent to which the reader's sensitivity to linguistic word properties is affected by the reading task. To this end, we analyze the strength of the response to frequency, predictability, and word length across the Gathering and Hunting conditions and interest areas. We quantify word predictability using surprisal (Hale, 2001; Levy, 2008), defined as $-\log_2(p(word|context))$, where *context* is the textual content in the paragraph preceding the *word*. The surprisal values are obtained from the GPT-2 language model (Radford et al., 2019; Wolf et al., 2020), whose surprisal values have been shown to correlate well with reading times (Wilcox et al., 2020; Heilbron et al., 2022; Shain et al., 2022). For frequency estimates we use unigram surprisal, defined as $-\log_2(p(word))$. We use word frequency counts from Wordfreq (Speer, 2022), which is based on multiple corpora.Word length is defined as

the number of characters in the word, excluding punctuation.

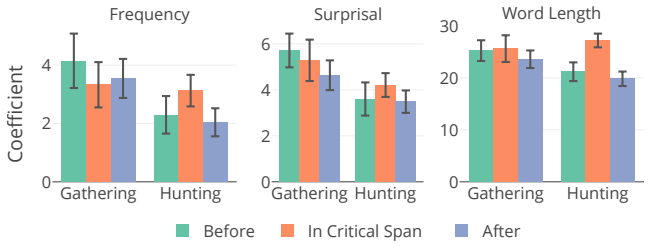

Figure 3: Frequency, surprisal, and word length effects on TF times. Depicted are current word coefficients from a linear mixed-effects model predicting TF times from these properties of the current and previous words fitted separately for each reading condition (Hunting, Gathering) and interest area (Before, In, After).[5] Error bars represent 95% confidence intervals.

We examine reading time response to word properties using the coefficients of a linear model which predicts TF times from current and previous word frequency, surprisal and word length in each combination of interest area and condition. We then examine the resulting model coefficients for the current word, depicted in Figure 3. We observe a pattern of results consistent with the TF times analysis in Figure 1. Crucially, differently from the Gathering condition where the responses to linguistic properties of the text are similar or lower within the Critical Span compared to the Before and After areas (all but one case), in the Hunting condition they are *higher* in the Critical Span than in the Before and After areas ($p < 10^{-4}$)[6] (with significant interactions across the reading conditions ($p < 10^{-8}$ in all cases but one). Critical span response is similar across the Hunting and Gathering conditions, while Before and After responses are lower in the Hunting condition compared to the Gathering condition ($p < 10^{-4}$). The SM presents this analysis for the earlier measures FF and GD (SM Figures 7 and 8 respectively), where again, in line with the TF results, word property effect differences across reading conditions and within interest area in the Hunting condition are attenuated or disappear. This provides further evidence that differences in text engagement across conditions and interest areas stem to a large degree from word re-reading.

Overall, these results suggest that the seek–solve–wrap-up strategy proposed in the context of TF times and regressions is also manifested in the degree to which reading times are influenced by the text. Readers are less engaged with the text in the Before and After Hunting areas compared to the Gathering condition and the Hunting Critical Span. The combination of these results with the TF and RR analyses suggests that the task influences reading patterns and text engagement in a systematic manner across all three interest areas.

Reading Comprehension Performance

In our final analysis we ask how eye movements in information-seeking and ordinary reading interact with reading comprehension performance. Thus far we characterized both regimes using all the experimental trials. However, in some cases participants do not perform the task successfully, and choose an incorrect answer to the question. The percentage of questions answered correctly is 81.9 in the Gathering condition and 86.9 in the Hunting condition ($p < 10^{-6}$). In Table 1 we further provide a breakdown of all the trials by reading condition and answer type (A–D), where the answer types are ordered by degree of comprehension.

	Gathering		Hunting	
A	5,878	(81.9)	6,379	(86.9)
B	723	(10.1)	556	(7.6)
C	391	(5.4)	295	(4.0)
D	188	(2.6)	114	(1.5)
Total	7,180	(100)	7,344	(100.0)

Table 1: Number of trials by condition (Hunting versus Gathering) and answer type. Values in parentheses are percentages.

We focus on the fundamental finding that differently from the Gathering condition, in the Hunting condition TF times are longer in the Critical Span than outside it. We examine the extent to which this behavior interacts with the degree of success in answering the question. To this end, in Figure 4 we present Hunting TF times within and outside the Critical Span by answer type for both reading conditions. Focusing on the Hunting condition, we first note a general trend whereby reading times in the Critical Span decrease along with the quality of the answer. At the same time, reading times outside the Critical Span increase for the answers that are not anchored in the Critical Span, and are highest for the *C* answer which has support outside of the Critical Span.

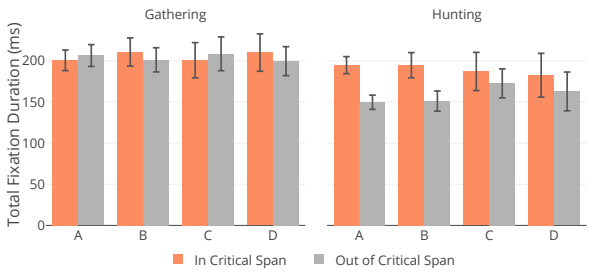

Figure 4: Total Fixation Duration by reading condition, chosen answer (A/B/C/D), and interest area (within versus outside the Critical Span). Error bars represent 95% confidence intervals.

Importantly, we find that for trials where participants choose either the correct answer *A*, or the distractor that represents a miscomprehension of the Critical Span (*B*), reading times are

[5] $TF \sim freq * len + surp + freq_{prev} + surp_{prev} + len_{prev} + (freq + surp + len|subj) + (freq + surp + len|parag)$. Random effects structure is simplified due to model convergence issues.

[6] Statistical tests are performed using a linear model: $TF \sim freq * len + freq * span + len * span + surp * span + freq_{prev} + surp_{prev} + len_{prev} + (span + freq + surp + len|subj) + (span + freq + surp + len|parag)$ for across-span tests and similarly with *condition* instead of *span* for across-condition tests.

longer within the Critical Span than outside it ($p < 10^{-34}$ for A and $p < 10^{-10}$ for B). However, in trials where participants chose the answer that has support outside the Critical Span (C) this difference is not statistically significant ($p = 0.18$). Similarly, when participants chose D, an answer that has no support in the paragraph, the difference is only marginally significant ($p = 0.049$). The TF differences between inside and outside of the Critical Span are smaller in C and in D when compared to the difference in the A and B trials ($p < 0.05$). In Gathering, we see no evidence for differences in inside versus outside the Critical Span TF times across answer types.

Following the results of this analysis, in Figure 5 we provide a split of Figure 1 to answers A/B and C/D in the Hunting condition, where we see that the pattern of slopes for C/D falls between Hunting A/B and the Gathering condition. The SM further contains the analyses from Figure 2 and Figure 3 by answer type (SM Figures 9 and 10 respectively), suggesting that the patterns observed in these analyses are similarly driven by A and B trials, and are attenuated in C and D trials.

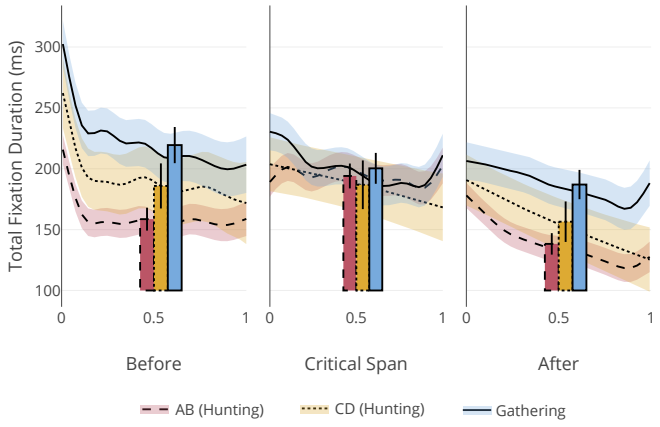

Figure 5: TF times analysis from Figure 1 with Hunting trials split to A/B and C/D answers.

Discussion

In this work we asked how information seeking reading behavior differs from ordinary reading.[7] To make progress on answering this question we leverage a dataset which contains both ordinary and task-based reading samples, textual annotations for task-critical information, and a structured reading comprehension component. Combining these sources of information we find that eye movements in information seeking regimes are (1) systematically different from ordinary reading, (2) optimized for the specific task at hand, and (3) interact with task performance.

[7]We use the term "ordinary reading" to refer to situations where the reader is expected to read attentively while not having a specific goal beyond general comprehension of the text. While we believe that such situations are common, following the discussion in Huettig & Ferreira (2022), we acknowledge that the term is somewhat problematic, especially when the experimental setup includes comprehension questions which are typically not part of our daily experience.

Our analyses provide a fine grained characterization of these differences through reading times, regressions, response to linguistic properties of the text, and reading comprehension performance. They suggest that in information seeking regimes the average reader engages with the text in a strategic and resource efficient manner, which can be broadly divided into three stages: seek, solve, and wrap-up. In the first stage, readers look for task-critical information while engaging in a reading behavior characterized by word position constant and short reading times and weak responses to linguistic properties of the text. Once task-critical information is identified, reading times remain constant, but with similar average reading times and responses to word properties to ordinary reading, stemming in large part from re-reading of task critical information. Once readers progress past task-critical information their reading times become again short and now also rapidly decreasing as the distance from task relevant information grows, with weak word property responses. However, even at this stage, task-critical information plays a central role in the reading dynamics by attracting frequent regressive eye movements. We further relate reading patterns to reading comprehension performance. This analysis reveals a correspondence between higher engagement with task-critical compared to non-critical information, and reading comprehension behavior. The worse the readers perform the task, the more attenuated are the information seeking effects we find.

Our results are consistent with a rational account of task-based reading. The seek–solve–wrap-up strategy allows readers to spend less cognitive effort (as reflected in reading times and word property effects) on task irrelevant information, while still leading to a higher reading comprehension performance than in ordinary reading. Importantly, standard computational models of eye movements in reading such as EZ-Reader (Reichle et al., 2003) and SWIFT (Engbert et al., 2005) do not directly model information seeking regimes. The NEAT model by Hahn & Keller (2023) addresses such regimes but predicts only word level reading times rather the entire sequence of fixations, and does not model the interplay with reading comprehension. Our work provides new empirical results that will inform the development of future fine-grained computational models for eye movements in reading in both ordinary and information seeking regimes.

Conclusion

This work sheds new light on the relationships between eye movements, task-based reading, and reading comprehension. Overall, we find that readers adjust their behavior to the given task in an efficient manner consistent with a rational account of cognitive resource allocation. We further find that reading behavior around task-critical information is correlated with question answering behavior. These results will inform future approaches for tracking cognitive state and reading comprehension in real time, and will further guide the development of computational models for eye movements in reading during both ordinary and task-driven reading.

Acknowledgments

This work was supported by ISF grant 2070358.

References

Berzak, Y., Malmaud, J., & Levy, R. (2020). STARC: Structured Annotations for Reading Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*. Retrieved 2021-10-26, from https://github.com/

Berzak, Y., Nakamura, C., Smith, A., Weng, E., Katz, B., Flynn, S., & Levy, R. (2022, apr). Celer: A 365-participant corpus of eye movements in l1 and l2 english reading. *Open Mind*, 1-10.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological review*, *112*(4), 777.

Hahn, M., & Keller, F. (2023). Modeling task effects in human reading with neural network-based attention. *Cognition*, *230*, 105289.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022, August). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. Retrieved 2022-10-10, from https://pnas.org/doi/full/10.1073/pnas.2201968119 doi: 10.1073/pnas.2201968119

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, *28*.

Hollenstein, N., Tröndle, M., Plomecka, M., Kiegeland, S., Özyurt, Y., Jäger, L. A., & Langer, N. (2021, December). Reading Task Classification Using EEG and Eye-Tracking Data. *arXiv:2112.06310 [cs]*. Retrieved 2022-05-12, from http://arxiv.org/abs/2112.06310 (arXiv: 2112.06310)

Hollenstein, N., Tröndle, M., Plomecka, M., Kiegeland, S., Özyurt, Y., Jäger, L. A., & Langer, N. (2022, March). *The ZuCo Benchmark on Cross-Subject Reading Task Classification with EEG and Eye-Tracking Data* (preprint). Neuroscience. Retrieved 2022-04-24, from http://biorxiv.org/lookup/doi/10.1101/2022.03.08.483414 doi: 10.1101/2022.03.08.483414

Huettig, F., & Ferreira, F. (2022). The myth of normal reading. *Perspectives on Psychological Science*, 17456916221127226.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Just, M. A., Carpenter, P. A., & Masson, M. (1982). What eye fixations tell us about speed reading and skimming. *Eye-lab Technical Report Carnegie-Mellon University, Pittsburgh*.

Kaakinen, J., & Hyönä, J. (2010, November). Task Effects on Eye Movements During Reading. *Journal of experimental psychology. Learning, memory, and cognition*, *36*, 1561–6. doi: 10.1037/a0020693

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2002). Perspective effects on online text processing. *Discourse processes*, *33*(2), 159–173.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004, January). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284. Retrieved 2022-06-22, from https://doi.org/10.1080/09541440340000213 (Publisher: Routledge _eprint: https://doi.org/10.1080/09541440340000213) doi: 10.1080/09541440340000213

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Malmaud, J., Levy, R., & Berzak, Y. (2020). Bridging Information-Seeking Human Gaze and Machine Reading Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 142–152). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved 2021-10-07, from https://www.aclweb.org/anthology/2020.conll-1.11 doi: 10.18653/v1/2020.conll-1.11

Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European journal of cognitive psychology*, *16*(1-2), 3–26.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

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.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 514–528. (Place: US Publisher: American Psychological Association) doi: 10.1037/a0020990

Rayner, K., Warren, T., Juhasz, B., & Liversedge, S. (2004, December). The Effect of Plausibility on Eye Movements in Reading. *Journal of experimental psychology. Learning, memory, and cognition*, *30*, 1290–301. doi: 10.1037/0278-7393.30.6.1290

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, *26*(4), 445.

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010, September). Eye movements during mindless reading. *Psychological Science*, *21*(9), 1300–1310. (Publisher: SAGE PublicationsSage CA: Los Angeles, CA) doi: 10.1177/0956797610378686

Rothkopf, E. Z., & Billington, M. (1979). Goal-guided learning from text: inferring a descriptive processing model from inspection times and eye movements. *Journal of educational psychology*, *71*(3), 310.

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, *131*(1), 1–27.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. P. (2022, Nov). *Large-scale evidence for logarithmic effects of word predictability on reading time.* PsyArXiv. Retrieved from psyarxiv.com/4hyna doi: 10.31234/osf.io/4hyna

Speer, R. (2022, September). *rspeer/wordfreq: v3.0.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7199437 doi: 10.5281/zenodo.7199437

Tokunaga, T., Nishikawa, H., & Iwakura, T. (2017, September). An eye-tracking study of named entity annotation. In *Proceedings of the international conference recent advances in natural language processing, RANLP 2017* (pp. 758–764). Varna, Bulgaria: INCOMA Ltd. Retrieved from https://doi.org/10.26615/978-954-452-049-6_097 doi: 10.26615/978-954-452-049-6_097

Tomanek, K., Hahn, U., Lohmann, S., & Ziegler, J. (2010, July). A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1158–1167). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P10-1118

Vajjala, S., & Lučić, I. (2018, June). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 297–304). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved 2022-04-24, from https://aclanthology.org/W18-0535 doi: 10.18653/v1/W18-0535

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-demos.6 doi: 10.18653/v1/2020.emnlp-demos.6

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673-686.

Wood, S. N., Goude, Y., & Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 139–155.