NORTHWESTERN UNIVERSITY

Word Identification and Eye Movement Control in Reading as Rational Decision Making

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

By

Yunyan Duan

EVANSTON, ILLINOIS

June 2021

ABSTRACT

Word Identification and Eye Movement Control in Reading as Rational Decision Making

Yunyan Duan

This dissertation provides evidence that reading is best explained as rational gathering of visual information to identify words efficiently. Although empirical evidence from human reading research suggests a close link between eye movements and cognitive process, it is not clear how readers decide when and where to move their eyes as a function of their cognitive states, and why they make certain eye movement decisions the way they do. The standard model of word identification assumes that a word requires a fixed amount of time to identify, which is a function of its word frequency, predictability, and the distance between the fixation location and the word center. Due to visual acuity constraints, reading time is minimized at the word center. Since word identification is considered the main driver of eye movements, dominant models of eye movements in reading assume that readers always target their eyes toward the word center to obtain best-quality visual information of the whole word and thus minimize the time needed to identify the word. In contrast, the rational model of eye movements in reading considers a word to be identified from a combination of visual and linguistic information, sensitive to

the interaction of these two sources of information and therefore sensitive to that word's particular visual neighborhood structure. As a result, readers move their eyes to the position that best distinguishes a word from its neighbors to identify the word quickly and accurately.

Previous modeling studies have shown that a rational model of eye movements in reading provides natural explanations for several eye movement phenomena, which can also be explained in dominant models but in less parsimonious ways. In this dissertation, we present qualitative evidence that the rational model explains eye movement phenomena that cannot be explained by dominant models, and quantitative evidence that the rational model better predicts eye movement phenomena. Specifically, in Chapter 2 we demonstrate that human readers seek visual information of the uncertain part of the word instead of always targeting the word center when they decide where to refixate, which cannot be explained by dominant models with a standard account of word identification. In Chapter 3 we demonstrate that the letter position that maximizes word identification efficiency varies as a function of the structure of the particular word, which is only predicted by a rational model. In Chapter 4, we provide quantitative evidence that the rational model predicts humans' decision to skip a word better than a model based on the standard account of word identification. In Chapter 5, we provide quantitative evidence that predicting readers' comprehension from eye movements through a rational model is more robust than through using the features from a dominant model.

Altogether, this dissertation provides evidence that the rational model of eye movements in reading, which captures the complicated interaction between visual and linguistic information and optimizes eye movement decisions accordingly, is able to better explain and predict human eye movements than dominant models of eye movements that take a standard account of word identification. Therefore, the eye-mind link between eye movements and online language processing is naturally understood as rational eye movement decision making based on the knowledge from a probabilistic word identification process.

Acknowledgements

First and foremost, I would like to thank my advisor, Klinton Bicknell, for his support and guidance throughout my graduate career. This dissertation would not be possible without his support. I feel so thankful for his insightful suggestions about stating the research questions clearly and sharpening the arguments, for his genius ideas about mathematical modeling, and for his patience reviewing my writings and helpful feedback, which really helps me develop my mind as a researcher. I feel lucky to study eye movements in reading under his guidance, which is exactly the kind of research that I am interested in and is in the direction I think where the field should go. Besides research, I also learned from him to responsible and positive in life. I will be forever grateful to you for the time you devoted to my mentorship.

I would like to thank Matt Goldrick and Diego Klabjan for serving on my committee and providing support and guidance. I especially would like to thank Matt for his helpful feedback for both my QP and for my dissertation, and for having me in his lab meeting, which keeps me emerged in linguistic research.

I would also like to thank other faculty and members of the linguistics department at Northwestern University for teaching and sharing knowledge of linguistics. Outside of Northwestern, I would like to thank Yevgeni Berzak and Roger Levy for their valuable advice on the study in Chapter 5 of this dissertation.

I am thankful for my former peers in Language and Computation Lab, in particular Adam Goodkind, Jordan Hosier, and Nicole Mirea, for our collaboration on projects and for your constructive feedback on my work in the early days. I am also thankful for members of SoundLab for your support and for your feedback on my presentations. I would also like to thank my cohorts, Jordan Hosier, Ethan Myers, Haley Farkas, and Wei Song, for taking courses and spending time together, and for helping me with my English. And I would like to thank alumni of the department for their help along the way. I would like to thank Tommy Denby for his help during coursework and during my internship. I would also like to thank Nayoun Kim for her warm help on every aspect of my life at Northwestern, and I enjoyed all our discussions, our travels to conferences together, and so on.

I would like to thank the English Language Program staff at Northwestern for the training and events they offer to help international students (like me) improve their English. I would also like to thank CCIS, and especially my dear language partner Doris Liberman for her help and friendship since I started at Northwestern.

I would like to thank Professor Xiaolin Zhou, my undergraduate advisor, for inspiring my interest in cognitive science. I would like to thank Drs. Shaorong Yan, Yingyi Luo, Hongbo Yu, Lihui Wang, Ciyang Qing, Wangshu Feng, Yingying Tan, and other members of CBCS at Peking University. Thank you for being models of what good researchers should be and what kind of person I would like to become. I would like to thank my collaborators, Drs. Yanan Sheng and Fuyun Wu, for working on linguistic projects together. I would also like to thank Dr. Ou Wu, who inspired my interest in machine learning and guided me through a real machine learning project.

Last but not least, I would like to thank my friends and family, for their love and support along this long journey.

This research was supported by a National Science Foundation award 1734217 to Klinton Bicknell.

Chapter 2 is expanded from Duan, Y., & Bicknell, K. (2017) [Refixations gather new visual information rationally. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 301-306.] The dissertation author was the primary investigator and author of this paper. In addition to being presented to the Cognitive Science Society, this work was also presented at the 22nd annual conference on Architecture and Mechanisms for Language Processing (AMLaP).

Chapter 4 in full, is an exact copy of the material as it appears in Duan, Y., & Bicknell, K. (2020) [A rational model of word skipping in reading: ideal integration of visual and linguistic information. *Topics in Cognitive Science*, 12(1), 387-401. doi:10.1111/tops.12485.] The dissertation author was the primary investigator and author of this paper. In addition to being presented to the Cognitive Science Society, this work was also presented at the 32nd annual CUNY Conference on Human Sentence Processing.

A section (Analysis 1) of Chapter 5 was presented as a poster at the 32nd annual CUNY Conference on Human Sentence Processing.

Table of Contents

ABSTRACT 3

Acknowledgements 6

Table of Contents 10

List of Tables 13

List of Figures 14

Chapter 1. Introduction 18

1.1. Eye movements in reading 21

1.2. Rational models of reading 22

1.3. Motivations and approaches 23

1.4. Overview of chapters 24

Chapter 2. Word Identification is Constructive: Refixation Seeks New Information[1] 30

2.1. Introduction 30

2.2. Experiment 1: Human data in Dundee corpus 35

2.3. Experiment 2: Human data in Potsdam Sentence Corpus 38

2.4. Experiment 3: E-Z Reader 41

2.5. Rational models of reading 44

2.6. Experiment 4: Rational model 49

2.7. General discussion 52

2.8. Acknowledgments 53

Chapter 3. A Rational Model of Within-word Eye Movements via Reinforcement Learning 54

3.1. Introduction 54

3.2. Within-word eye movements 59

3.3. General assumptions and problem statement 63

3.4. Model implementation 69

3.5. Experiment 71

3.6. Results 76

3.7. Discussion 82

Chapter 4. A Rational Model of Word Skipping in Reading: Ideal Integration of Visual and Linguistic Information[2] 87

4.1. Introduction 87

4.2. Related work 92

4.3. Rational model of skipping 95

4.4. Experiment 98

4.5. Results 101

4.6. Discussion 104

4.7. Acknowledgments 106

Chapter 5. Inferring Sentence Comprehension from Eye Movements in Reading 107

5.1. Introduction 107

5.2. Experiment and data collection 114

5.3. Analysis 1: Eye movement features predictive of comprehension 116

5.4. Analysis 2: Using machine learning to predict comprehension for new trials 119

5.5. Analysis 3: Rational model simulation 124

5.6. General discussion 128

Chapter 6. Conclusion 132

6.1. Summary of results 132

6.2. Future directions 133

References 135

References 135

List of Tables

3.1 Hyperparameters of the *gym-wordreading* environment. 68

3.2 Properties of “beginning” and “end” words. 76

4.1 Generalized additive mixed-effects regression model results of the baseline model (note that random slopes for these fixed effects were not included in the model; the model included a random intercept over participants). The GAMM was fitted by REML, and p-values were reported using *summary.gam* function in *mgcv* package (Wood, 2011). 102

4.2 Significance of averaged entropy of a rational model’s posterior distribution when added to the baseline model. 103

5.1 Example sentence. The high-frequency neighbor (HFN) of the Target word (**minuet**) is **minute**. 115

5.2 Generalized linear mixed-effects model results of eye movement features and word frequency features. 120

List of Figures

2.1 The standard model and the rational model make different predictions for where refixations go. For the standard model, refixations always target the center of the word, regardless of launch site. For the rational model, refixations target positions where character identity has low confidence (here represented by hypothetical $m(j)$ values). Therefore, closer launch sites, which provide more visual information about the word's initial letters (schematically represented here by grey rectangle) predict refixations are more likely to move forward. The refixation decisions here are based on eye movement policy parameters of $\alpha = .9$ and $\beta = .7$. (See Eye movement policy section for more details.) 35

2.2 Effect of launch site on proportion of forward-moving refixations on data from Dundee corpus. Each panel contains data from a combination of word length and landing position, and shows a GAM smoother. 38

2.3 GAMM estimation of effect of launch site on proportion of forward-moving refixations on data from Dundee corpus (left panel) and PSC corpus (right panel) with initial fixations landing on left-half of the word. 39

2.4 Effect of launch site on proportion of forward-moving refixations on data from PSC corpus. Each panel contains data from a combination of word length and landing position, and shows a GAM smoother. 42

2.5 Effect of launch site on proportion of forward-moving refixations in data from E-Z Reader simulation. Each panel contains data from a combination of word length and landing position and shows a GAM smoother. 45

2.6 GAM estimation of effect of launch site on proportion of forward-moving refixations on data from E-Z Reader (left panel) and rational model (right panel) simulation with initial fixations landing on left-half of the word. 46

2.7 Effect of launch site on proportion of forward-moving refixations in data from rational model simulation. Each panel contains data from a combination of word length and landing position and shows a GAM smoother. 51

3.1 An example of the procedure of a trial with true word being ‘million’ and initial landing position set to 6. 68

3.2 Structure of the actor-critic object using a CNN model. 72

3.3 Average reward per epoch of training with a CNN model. 78

3.4 Average evaluation reward of different policies in a ‘speed’ setting. 79

3.5 Overall gaze duration and refixation rate as functions of initial landing position in ”speed” and ”accuracy” settings. 80

3.6 Word frequency effect as a function of initial landing position in terms of gaze duration and refixation rate in ”speed” and ”accuracy” settings. 81

3.7 Gaze duration and refixation rate as functions of initial landing position and word frequency in human data. Plots are adapted from O’Regan, 1992, Figure 20.2. Note that only fixation positions 1, 2.5, 4, 5.5, and 7 were plotted. 82

3.8 Gaze duration and refixation of “beginning” vs. “end” words in ”speed” and ”accuracy” settings. 83

3.9 Gaze duration for “beginning” and “end” words in human data. Plots are adapted from O’Regan, 1992, Figure 20.3. Note that the data were from a task of identifying words with the length of 10 or 11, and only fixation positions 2, 4, 6, 8, and 11 were plotted. 84

4.1 Partial effect of postH with a frequency prior in predicting skipping rate. 103

4.2 Partial effect of postH with a 5-gram prior in predicting skipping rate. 104

5.1 Illustration of the data split for new subject and new item regimes. 123

5.2 AUCs of logistic regression (LR) and XGBoost (XGB) models with text, eye movement, and all features to predict correct/incorrect word identification. 125

5.3 AUCs of logistic regression (LR) and XGBoost (XGB) models with rational model's prediction of the target word's probability to predict correct/incorrect word identification. 129

CHAPTER 1

Introduction

During reading, the visual processing of language, readers are able to identify text and extract meaning from visual input. An important component of this process is eye movements, as readers must move their eyes constantly to bring different portions of the text into a small area of the retina with high visual acuity. In natural reading, readers make a quick eye movement (called a saccade) to send the eyes to a location, remain relatively still there for about 200-300 ms to gather visual input around that region (called a fixation), and launch the next saccade. This process can be captured by the eye-tracking technique, which records the eyes' locations in real-time. Psycholinguistic researchers use eye-tracking to study online human language processing, given that long reading time and multiple fixations on a word often indicate processing difficulty related to this word.

Ample evidence from empirical studies have found that eye movements are sensitive to word-level linguistic information, such as word frequency (Schilling, Rayner, & Chumbley, 1998; Kliegl, Grabner, Rolfs, & Engbert, 2004; McConkie, Kerr, Reddix, Zola, & Jacobs, 1989) and contextual predictability (Balota, Pollatsek, & Rayner, 1985; Rayner & Well, 1996); and visual information, such as word length (McConkie, Kerr, Reddix, & Zola, 1988; Brysbaert & Vitu, 1998) and preview obtained from preceding fixations (Inhoff, Eiter, Radach, & Juhasz, 2003; Angele & Rayner, 2013). Researchers therefore believe that there is a strong link between real-time language processing and eye movements (Rayner, 1998; Morrison, 1984; Just & Carpenter, 1980). A standard way of modeling word identification in reading is thus assuming that each word is associated with some 'word difficulty', which is a function of word frequency and contextual predictability. Readers have to fixate a word for a certain amount of time, which is a function of both word difficulty as well as the distance between fixation position and the word center. Readers, therefore, must be in one of two states: they either do not identify the word at all at any time point before reaching the required amount of time; or they 'identify' the current word completely once they fixate (or, focus their attention on) the word for long enough. Dominant models of eye movements in reading adopt the standard model of word identification, and make heuristic eye movement decisions depending on the state. They estimate the parameters of the reading time function (and many other functions) by fitting human eye movements data. These computational models can yield eye movement trajectories similar to humans' (Reichle, Warren, & McConnell, 2009; Engbert, Nuthmann, Richter, & Kliegl, 2005).

Although this line of modeling work suggests that both visual and linguistic information play important roles when making eye movement decisions, it does not explain the mechanism of eye movements in reading beyond mathematical functions with estimated parameters. As a result of the fixed function forms and parameters, these models could only make predictions at a restricted scope, and are not expected to accommodate eye movement phenomena affected by finer-grained information (e.g. a word's particular visual neighborhood structure). To explain *how* the visual and linguistic information is processed and *why* readers move their eyes the way they do, more research is needed beyond the description of eye movement behaviors.

In contrast, the rational model of reading provides insights into how eye movements are generated from processing various information available during reading (Legge, Klitz, & Tjan, 1997; Legge, Hooven, Klitz, Mansfield, & Tjan, 2002; Bicknell & Levy, 2010, 2012b). The rational model of reading is based on the idea of rational analysis (Anderson, 1990), which tries to explain cognition in terms of an agent's optimal adaptation to the environment. The philosophy of this approach is that researchers can understand why humans behave in certain ways by comparing human behaviors with the optimal behaviors yielded by an agent, where the agent performs the task with an environment model that specifies the information available to the agent and the constraints the agent faces. If the agent's behaviors are similar to that of humans, then it indicates that humans perform near-optimally under the given environment. This approach therefore both explains the motivation of human behaviors, and also specifies the conditions and constraints that humans face.

Previous rational models of eye movements in reading have shown that several human eye movement phenomena can be explained from a rational perspective (Legge et al., 1997, 2002; Bicknell & Levy, 2010, 2012b). This dissertation furthers this line of research by evaluating the rational framework with more eye movement phenomena, especially those predicted differently by a rational model and eye movement models with the standard account of word identification. Before presenting the specific questions addressed by the dissertation, we first summarize the empirical findings of eye movements and existing rational models of eye movements in reading.

1.1. Eye movements in reading

To understand how reading works, researchers have developed several paradigms. One is natural reading, in which participants read a piece of text (usually a sentence) while their eye movements are recorded. This is similar to reading in our daily life, and is more natural than other psycholinguistic tasks – for example, self-paced reading, in which participants read one word/phrase at a time, push a button to move to the next word/phrase, and cannot go back to the previous text. Natural reading is probably the most commonly used task in the research of reading. Researchers use natural reading to collect eye movement data from a large number of participants as they read a long piece of text, and create eye movement corpora. The most known eye movement corpora include Schiling corpus (Schilling et al., 1998), Dundee corpus (Kennedy, Hill, & Pynte, 2003; Kennedy & Pynte, 2005), Potsdam Sentence Corpus (Kliegl, Nuthmann, & Engbert, 2006), and GECO (Cop, Dirix, Drieghe, & Duyck, 2017), among many others.

Eye-tracking is also used to study how an isolated word is identified. In this task, one word is first presented somewhere in the participant's visual field and the participant then moves their eyes freely to identify the word. Compared to natural reading, this task excludes influence from context, and allows researchers to control where the participant fixates at the very beginning. This task yields more fine-grained observations regarding humans' eye movements on a single word (O'Regan, Lévy-Schoen, Pynte, & Brugaillère, 1984; O'Regan & Lévy-Schoen, 1987; O'Regan, 1992). Most effects observed in single word id are also observed in natural reading, though often somewhat attenuated due to context and preview (Rayner & Well, 1996).

Through eye-tracking experiments using natural reading, isolated word identification, and other paradigms, researchers have reached the consensus that word identification can be considered as a primary driving force of eye movements. They also agree that both visual and linguistic factors matter for deciding when and where to move the eyes. With this in mind, any model of eye movements in reading should be able to reproduce typical eye movement patterns found in human reading experiments.

1.2. Rational models of reading

Rational models of reading frame the reading behavior as a process in which readers move eyes to seek information that maximizes word identification efficiency (Bicknell & Levy, 2010, 2012b; Legge et al., 1997, 2002). Specifically, readers combine visual and linguistic information to make a guess about the text identity. One of the earliest models is Mr. Chips (Legge et al., 1997, 2002), which models visual input as veridical characters from a window around the fixation, and models linguistic input as word frequency. Eye movements target the position that minimizes expected entropy about the word. Later rational models use a more realistic word identification model, which suggests that visual input is noisy rather than veridical. Specifically, they consider word identification to be a Bayesian belief updating process, in which a prior distribution over possible identities of the word given by the language model is combined with a likelihood term given by 'noisy' visual input conditional on the fixation position to form a posterior distribution over the identity of the word (Bicknell & Levy, 2010, 2012b). Eye movements are performed to gather visual information from a position where the reader is uncertain, if the probability of the most likely character at that position falls below some predefined confidence thresholds.

These models provide a unified explanation of the mechanism underlying eye movements, which is the rational gathering of visual information for word identification. They also yielded eye movement patterns similar to human eye movements, such as distribution of initial landing position on a word (Legge et al., 1997), word length effect (Legge et al., 1997, 2002), between-word regressions (Bicknell & Levy, 2010), and word frequency and predictability effects as reflected in several eye movement measures (Bicknell & Levy, 2012b).

1.3. Motivations and approaches

Although the rational model of eye movements in reading provides a natural way to explain several eye movement phenomena, these phenomena can potentially be explained by eye movement model with a standard account of word identification, though through more complicated ways. To better distinguish the rational model from dominant eye movement models that involve a standard account of word identification, we take two approaches: one is to examine phenomena for which the rational theory provides qualitative contrast with dominant models of eye movements, and the other is to quantitatively show that the rational model better predicts human eye movements than dominant models do.

In general, we obtain a rational model's predictions by implementing a computational model and running simulations. Similar to prior work of rational models of eye movements in reading, we consider word identification as Bayesian belief updating (Bicknell &

Levy, 2010, 2012b). Different from prior work, we focus on the identification of a single word, instead of a full sentence. This is advantageous because the computational cost is greatly reduced, allowing researchers to explore various representations of visual and linguistic information and different eye movement policies. Based on the result of word identification, namely the posterior distribution, we either get the rational model's eye movement strategy and qualitatively compare its behaviors to humans', or get a metric from the posterior distribution, and quantitatively evaluate its ability to predict eye movement phenomena and see if its prediction is more accurate than dominant models' prediction.

In this dissertation, we develop, implement, and evaluate four sets of computational simulations with the rational model, each of which is focused on a specific aspect of reading for which either the rational model makes qualitatively different predictions than models of eye movements that considers word identification in the standard way (namely, the direction of refixation as a function of the launch site in Chapter 2, and within-word eye movements as a function of specific word properties in Chapter 3), or the rational model is expected to provide better quantitative predictions (namely, skipping behaviors in Chapter 4, and readers' comprehension for sentences containing critical words with high frequency neighbors in Chapter 5).

1.4. Overview of chapters

The body chapters of this dissertation consist of four studies that evaluate rational models of eye movements at a fine-grained level. With Chapters 2 and 4 from published papers and Chapters 3 and 5 being unpublished work we plan to submit for publication soon, each chapter can stand on its own and can be read independently of the others. In this section, we briefly outline each chapter.

Portions of Chapter 2 have been published as Duan, Y., & Bicknell, K. (2017) [Refixations gather new visual information rationally. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 301-306.] Experiment 2 is the only section not included in the published paper. In this study, we examine where refixations (multiple fixations made on a word during first-pass reading) go as a function of the launch site. Specifically, in dominant models of eye movements in reading, visual information is processed holistically and a word is identified most efficiently when centered in the visual field. In contrast, rational models consider reading as a process maximizing word identification efficiency, in which visual information obtained from a series of eye fixations is processed constructively and is then synthesized to the language discourse. These two models yield different predictions regarding where the refixations go. The dominant model predicts that refixations always target the word's center, while rational models predict differently: As refixations depend on previous fixations, more backward movements to the word's beginning are expected if less information of this part has been obtained prior to fixating the word – when the prior fixation was further from the word.

We analyze the direction of refixations as a function of the launch site in two human eye movements corpora, one in English (Dundee corpus) and one in German (Potsdam Sentence Corpus), and find that both corpora show the consistent pattern that the refixations are more likely to go rightward for closer launch sites, as predicted by the rational account. We confirm that this pattern is not observed from a dominant model of eye movements (E-Z Reader) through simulation. We then implement a rational model of refixation, with word identification combining visual and linguistic information using Bayesian inference and eye movement decisions based on the confidence of letter identity at each position computed from the posterior, and find that this implementation of the rational account can indeed reproduce the human pattern. These results suggest that the effects of the launch site on where refixations go are inconsistent with models in which all intentional refixations target the word center, but are naturally yielded by a rational account of eye movements in reading.

Beyond the influence of visual information already obtained about a word on where refixations should go as we examine in Chapter 2, Chapter 3 evaluates more predictions of the rational account. In this chapter, we further examine the influence of both visual and linguistic information on both the 'when' and 'where' decisions of eye movements when identifying a word. Traditionally, eye movement research consider that separate pathways involve in deciding when and where to move eyes, leading to the prediction that eye movements target the word center as it minimizes effort needed to identify a word (in terms of gaze duration and refixation rate). In contrast, a rational model of reading considers that word identification incorporates both visual and linguistic information interactively and that eye movements follow an optimal strategy to maximize word identification efficiency, leading to the prediction that the position that minimizes reading effort is not only an additive function of initial fixation position and frequency, but also dependent on the structure of a word. For example, the positions to minimize the identification efforts are expected to shift towards word beginning for words with a rare beginning, and shift towards word end for words with a rare end.

We implement a rational model of eye movements in reading with a policy learned through deep reinforcement learning, and evaluate its behaviors by comparing them to human within-word eye movements. We find that the policy learned through reinforcement learning robustly outperforms heuristic policies by achieving higher confidence in a shorter time, and produces human-like behaviors in terms of overall effects of initial landing position, word frequency effects, and different reading time patterns for words with different structures. These results suggest that a rational model of eye movements in reading is able to explain within-word eye movements as resulting from rational combinations of visual and linguistic information and following optimal eye movement policy to maximize reading efficiency, whereas dominant models of eye movements not only requires extra assumptions about word identification and eye movement policy, but also fails to predict different reading time patterns for words with different lexical structures.

Chapter 4 has been published as Duan, Y., & Bicknell, K. (2020) [A rational model of word skipping in reading: ideal integration of visual and linguistic information. *Topics in Cognitive Science*, 12(1), 387-401. doi:10.1111/tops.12485.] In this study, we examine how eye movement decisions of skipping a word can be better predicted from the perspective of a rational model of eye movements in reading, which takes complicated interactions between visual and linguistic information into account. As observed in empirical studies, readers intentionally do not fixate some words, thought to be those they have already identified. In a rational model of reading, these word skipping decisions should be complex functions of the particular word, linguistic context, and visual information available. In contrast, dominant models of eye movements in reading only predict additive effects of word and context features. Here we test these predictions by implementing a rational model with Bayesian inference and predicting human skipping with the entropy of this model's posterior distribution. Results showed a significant effect of the entropy in predicting skipping above a strong baseline model including word and context features. This pattern held for entropy measures from rational models with a frequency prior, though not from models with a 5-gram prior. These results suggest complex interactions between visual input and linguistic knowledge as predicted by the rational model of reading, and that taking this into account provides better predictions for human skipping decisions better than dominant models of eye movements in reading.

In Chapter 5, we extend previous chapters' work of modeling eye movements for word identification into an application of using a rational model of eye movements to predict the outcome of reading a sentence. Specifically, we examine to what extent could comprehension be predicted from eye movements and whether eye movements predict comprehension in the way that rational models of reading expect by looking into the identification of words with a high frequency neighbor (HFN) during sentence reading. We directly test the rational model's predictions by examining machine learning models' performance with a feature generated from a rational model of reading, namely the (logit-transformed) probability of the target word. Results suggest that comprehension can be better predicted with an integrated metric generated by the Bayesian belief updating model of reading than with eye movement features alone. These results provide supportive evidence for the perspective of considering eye movements as rational behaviors of gathering visual information for text identification.

Chapter 6 summarizes the findings and how one can use the models proposed in this dissertation to further the understanding about rational eye movement decisions during reading.

CHAPTER 2

Word Identification is Constructive: Refixation Seeks New Information[1]

2.1. Introduction

Reading is a complex information processing task with a goal usually related to comprehending the text. In general, accurate text comprehension requires the identification of many (if not most) of the words in a text. It is not surprising, then, that decades of research on eye movements in reading have established that word identification can be seen as the primary driver of eye movements (Rayner, 1998). A substantial body of work has studied the role in this process of many information sources relevant to word identification in reading, including especially word frequency and in-context predictability, among others. However, although visual information is the primary source of information used to ultimately identify a word, the fundamental way in which visual information is used in word identification remains unresolved.

In the standard model of word identification in reading, word identification is hypothesized to be a holistic process, during which visual information about the word as a whole constrains the efficiency of identification. Eye movement studies have shown that a word presented in isolation is most rapidly identified when fixating approximately at its

[1] Portions of this chapter have been published in Duan, Y., & Bicknell, K. (2017). Refixations gather new visual information rationally. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 301-306.

center (O'Regan, 1990, 1992). It has also been found in natural reading that the fixation position that minimizes gaze duration (the total amount of time spent fixating a word in first pass) and refixation probability (the probability of making more than one fixation on a word in first pass) is on average at or slightly left of the center (Rayner, Sereno, & Raney, 1996). One explanation for these results is that when the word center is directly fixated, the largest possible part of the word falls in the central high-acuity portion of the visual field (the fovea), yielding the highest-quality visual input of the whole word; as the fixation deviates from the center, more letters of the word fall out of the fovea and suffer from a rapid drop in acuity, leading to poorer visual information about the overall word. Following this interpretation, it is hypothesized that visual processing efficiency of a word is maximized when fixating at word center, and decreases with increasing distance between word center and fixation position. This standard holistic account is incorporated in dominant eye movement models of eye movement control in reading (e.g. E-Z Reader, Reichle et al., 2009; and SWIFT, Engbert et al., 2005).

Alternatively, word identification may not utilize visual information holistically, especially in natural reading. Unlike in isolated word identification where information about a word comes only from visual input obtained by directly fixating it, in natural reading information about a word comes from more sources. These include contextual information from the preceding text and visual information obtained from fixations close to but not on the current word, which may still yield some visual preview of the word's initial letters. As a result, the most efficient positions from which to obtain useful new visual information about the word can vary from trial to trial, dependent on the information already obtained. Even in such an account, it is still possible that, on average, the most efficient positions are located near the center (as has been found in prior work). This account of word identification is implemented in rational models, which consider reading as a process of combining information from various sources to identify words and making eye movement decisions to maximize identification efficiency (Bicknell & Levy, 2010, 2012b; Legge et al., 1997, 2002). For example, if a reader in this framework is working to identify a particular word, considering all the information that has already been gathered, there may be parts of the word that the reader has already identified relatively well and parts that are still relatively uncertain. It is intuitive in such a situation that identification efficiency will be maximized by moving the eyes next to the part of the word about which the reader is still relatively uncertain. This is because such fixations would obtain fine-grained visual information of a particular part of the word, which can be combined with visual information obtained from previous fixations (as well as linguistic contextual information), and identify the word in a constructive manner. Thus, contrary to the holistic account's view that any fixation landing on a non-central position slows identification efficiency relative to a central fixation, the view from rational models is that the position in the word to move the eyes next to maximize identification efficiency will vary from trial to trial and depend on information already obtained.

A phenomenon that can be used to tease apart these two accounts is that of refixations, cases in which a word is fixated more than once during first-pass reading. The goal of an intended refixation is assumed to be moving the eyes to a position that will maximize identification efficiency of the current word. Despite previous experiments showing that refixation rate varied on average as a quadratic function of the distance between word center and the fixation position (McConkie et al., 1989) and was influenced by linguistic properties such as word frequency (Rayner et al., 1996), few studies shed light on where refixations go. The two accounts of word identification make different predictions for this question. The rational model predicts that refixations target the part of the word about which sufficient information has not yet been obtained. Which part of the word this is depends on the visual information already available.[2] In contrast, the standard model of word identification predicts that refixations should always target the center to maximize the holistic visual processing efficiency of the word, independent of information obtained about different parts of the word.

Naively, then, we could tease apart these two hypotheses by analyzing the relationship between the position of the initial fixation on a word (the 'landing position') and the refixation position. The rational account would predict that if the landing position is at the beginning of the word, a refixation should be at the end, and vice versa, whereas the standard model would predict that all refixations cluster around the center, regardless of the landing position. Empirically, this prediction of the rational models is borne out (Rayner et al., 1996), but the standard model explains this phenomenon in a different way. Specifically, there is a concept of *systematic error* (McConkie et al., 1988), which suggests that intended saccade sizes become biased toward the overall average saccade size. This means that refixation saccades intended to be short and target the center of the word in the standard model will tend to overshoot their target, landing on the opposite end of the word. Thus, both the standard model combined with systematic error and the rational model predict the effect of landing position on where refixations go.

[2] In general, the most efficient place to move the eyes next in a rational model depends not just on visual information already obtained but also contextual information. For the present paper, we ignore contextual information for simplicity.

Analyzing where refixations go as a function of the location of the *previous* fixation made before fixating a word (the ‘launch site’), however, can tease apart these two accounts, when controlling for effects of landing site. If a reader’s first eye movement to the word is launched from a position close to the word, then more visual information about the word’s beginning should be available (relative to the launch site being further away), holding constant the landing site. Therefore, rational models predict that for closer launch sites, a refixation should be less likely to move the eyes back toward the beginning of the word (Fig. 2.1, right panel). In contrast, the standard model would not predict such an effect, but predict that an intentional refixation that follows a fixation on the left half of a word should always go forward, while one that follows a fixation on the right half should always go backward, always targeting the word center (Fig. 2.1, left panel).

In this paper, we empirically evaluate these two competing predictions by performing a statistical analysis of where refixations go in a large eye movement corpus, and we compare these results to simulations from computational models of both accounts. In the next section, we report the results of our statistical analysis of human refixation data, showing that it is as predicted by the rational account. We then confirm that an eye movement model that implements the standard model cannot accommodate this finding by performing simulations with E-Z Reader (Reichle et al., 2009). After that, we describe our rational model of refixations. Finally, we confirm that simulations using it show the same qualitative pattern as the human data, and then conclude.

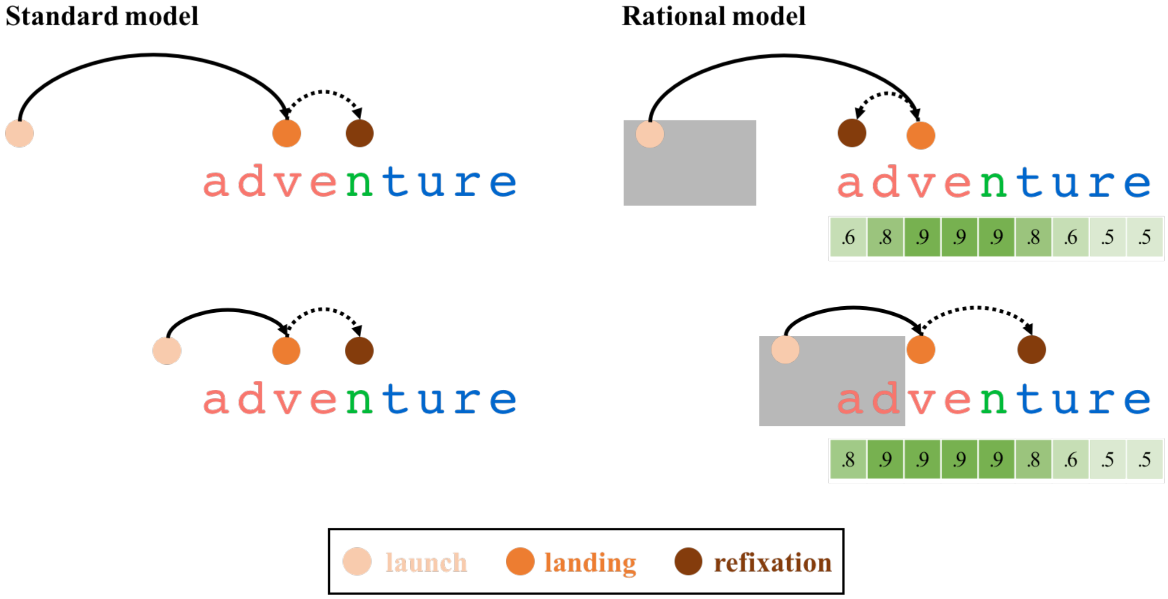

Figure 2.1. The standard model and the rational model make different predictions for where refixations go. For the standard model, refixations always target the center of the word, regardless of launch site. For the rational model, refixations target positions where character identity has low confidence (here represented by hypothetical $m(j)$ values). Therefore, closer launch sites, which provide more visual information about the word's initial letters (schematically represented here by grey rectangle) predict refixations are more likely to move forward. The refixation decisions here are based on eye movement policy parameters of $\alpha = .9$ and $\beta = .7$. (See Eye movement policy section for more details.)

2.2. Experiment 1: Human data in Dundee corpus

This analysis aims to tease apart the predictions of the rational model and the standard model on where refixations go. Specifically, we use the English part of the Dundee corpus (Kennedy, 2003) of eye movements during natural reading, and analyze the direction of refixation as a function of launch site, statistically controlling for landing site.

Methods

Data. The English part of the Dundee corpus contained eye movement records from 10 native English-speaking participants as they read through newspaper editorials (see Kennedy & Pynte, 2005 for further details.) We first did a set of screening procedures, according to criteria that are generally applied to eye movement data, to remove fixations involving blinks, non-first-pass fixations, and the first/last two fixations of the line. After this procedure, the corpus contained 23,854 fixations that were followed by a refixation during first-pass reading (18.9% of first-pass fixations). These data then underwent screening procedures excluding: (a) extremely far launch sites (1%), leaving the launch sites of fixations in the range $[-16, -1]$ (in terms of number of characters from word beginning); (b) fixations that landed on the space right before the word (25.5%) or on the last character of the word (4.7%) to ensure the variability of refixation directions; and (c) fixations on words of which the previous word was skipped to eliminate possible overshootings of the previous word (20.9%), since these can be followed by corrective saccades. In the end, the data consisted of 7,667 fixations.

Statistical analysis. A logistic generalized linear mixed-effects model (GLMM) was used to analyze the direction of refixations (forward vs. backward). Fixed effects included launch site and combinations of word length and landing site, which controlled for arbitrary effects of word length and landing site on refixation direction. Random effects included a random intercept and slope of launch site by subjects. Significance testing was via likelihood ratio test. All statistical analyses were implemented in the R environment, using the *glmer* function from the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) for GLMM implementation. In order to ensure model convergence, word length–landing site pairs for which all refixations (or all but 1) moved in the same direction were excluded, leaving 6714 fixations (87.6%).

Results and discussion

Fig. 2.2 shows the effect of launch site on the probability that refixations move forward for each word length–landing site pair. The GLMM showed that nearer launch sites predicted significantly more forward refixations, $\hat{\beta} = 0.15$, $SE = 0.03$, $\chi^2_1 = 13.98$, $p < 0.001$, 95% confidence interval $(CI) = [0.10, 0.20]$. As reported in the following section, the standard model can accommodate this effect only for landing sites on the right half of the word. To see whether this was also true of the human data, separate analyses were carried out for fixations with landing sites on the left and the right half of the word. For the left half (4790 fixations), launch site predicted more forward refixations, $\hat{\beta} = 0.16$, $SE = 0.03$, $\chi^2_1 = 10.91$, $p < 0.001$, $95\% CI = [0.09, 0.22]$, and the same was true for the right half (1362 fixations), $\hat{\beta} = 0.14$, $SE = 0.04$, $\chi^2_1 = 7.40$, $p < 0.01$, $95\% CI = [0.05, 0.22]$. As shown in Fig. 2.3 (left panel), which is plotted with all data of left half aggregated, the estimation from a GAMM on the probability of forward-moving refixations increases as launch sites get close to the word. These observations that closer launch sites predicted more forward-moving refixations confirm the rational model's predictions. The separate analyses of fixations on the left and right halves of the word indicated that this effect generalized across both.

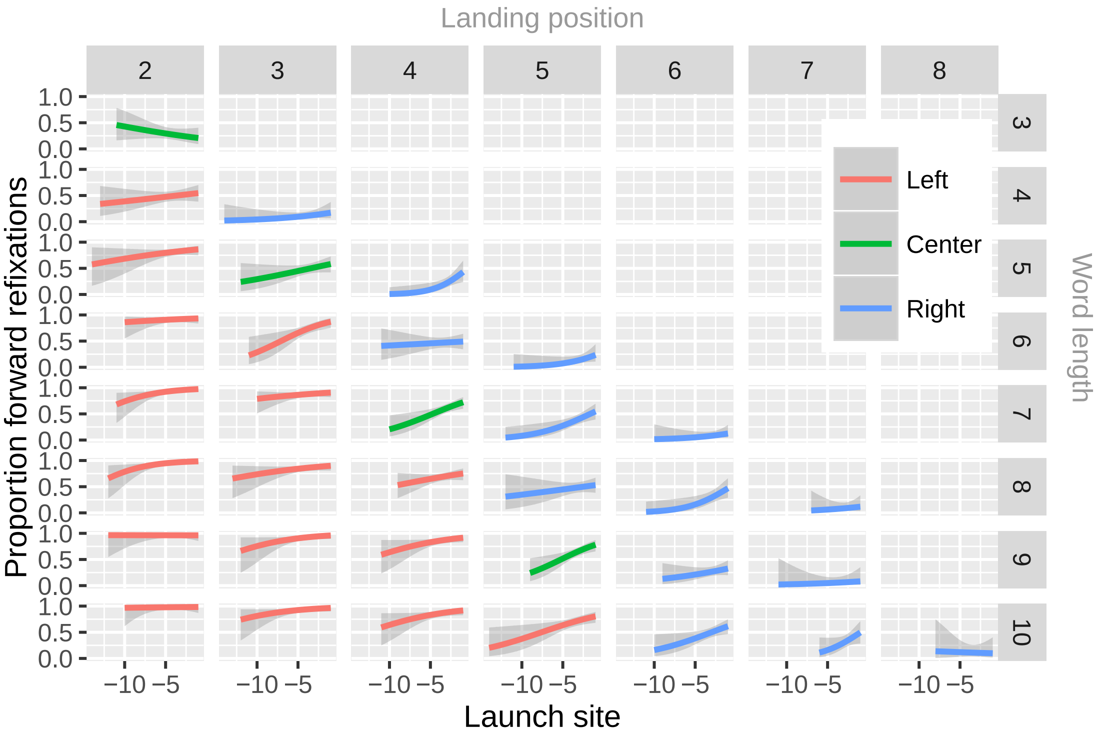

Figure 2.2. Effect of launch site on proportion of forward-moving refixations on data from Dundee corpus. Each panel contains data from a combination of word length and landing position, and shows a GAM smoother.

2.3. Experiment 2: Human data in Potsdam Sentence Corpus

To further confirm that the effect of launch site on the direction of refixations we observed in Expt. 1 is robust and does not result from specificity of the English language and/or the reading material, we examine the effect on a different eye movement corpus. Specifically, we carry out the same analysis as in Expt. 1 on the Potsdam Sentence Corpus of eye movements (PSC). This corpus differs from the Dundee corpus in language, type of text, and number of participants: 1) PSC is reading data of German while Dundee is reading data of English; 2) in PSC readers read single sentences from reading experiments

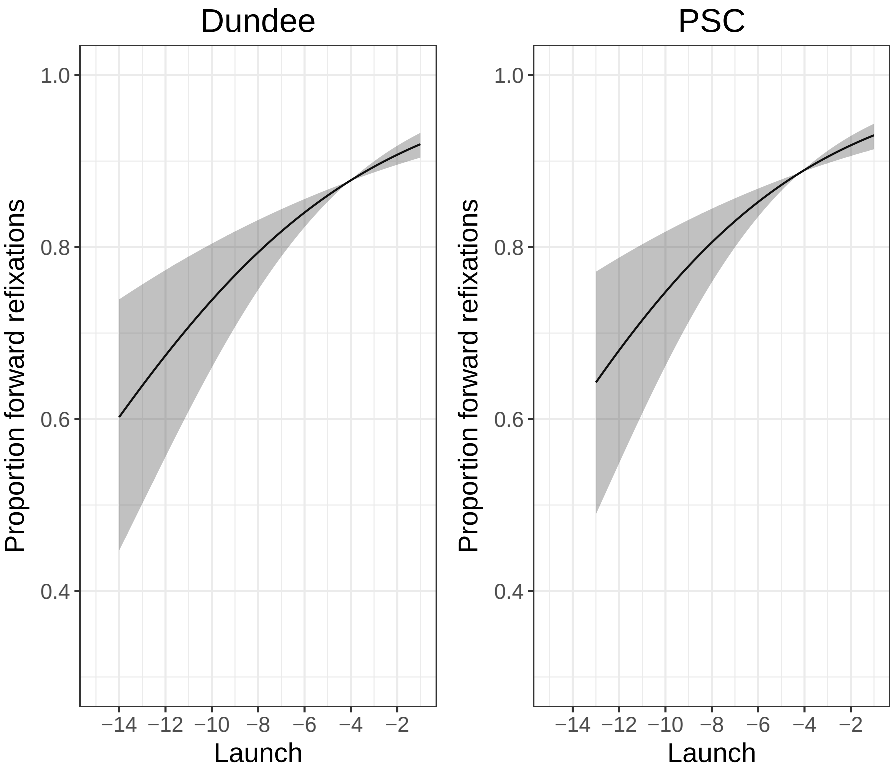

Figure 2.3. GAMM estimation of effect of launch site on proportion of forward-moving refixations on data from Dundee corpus (left panel) and PSC corpus (right panel) with initial fixations landing on left-half of the word.

while in Dundee corpus readers read continuous text from a newspaper; 3) PSC consists of a large number of participants (273 readers) reading short sentences while Dundee corpus consists of a few participants (10 readers) reading long texts. Observing the same effect on PSC as we found in the Dundee corpus – closer launch sites predict more forward fixations – will provide more evidence in favor of the rational model of word identification in reading.

Methods

Data. The dataset contained eye tracking data from 275 German-speaking participants as they read the Potsdam Sentence Corpus, which consisted of 144 single German sentences with a large variety of grammatical structures around a set of target words (See Kliegl et al., 2006 for further details about the corpus.)

The data cleansing procedure was the same as that in Expt. 1. Only the fixations followed by a refixation were included, which consisted of 13,940 fixations. These data then underwent screening procedures excluding: (a) fixations with launch sites higher than 99% of all launch sites, leaving the launch sites in the range [-13,-1] (in terms of number of characters from word beginning); (b) fixations that landed on the space right before the word (27.8%) or on the last character of the word (7.3%) to ensure variability of refixation directions; and (c) fixations on words of which the previous word was skipped to eliminate possible overshootings of the previous word (17.3%). In the end, the data consisted of 6,505 fixations.

Statistical analysis. A GLMM and a GAMM with the same fixed and random effects as that in Expt. 1 was adopted to analyze the effect of launch sites on refixation direction. Excluding word length-landing position pairs where all refixations (or all but 1) moved in the same direction left 4,220 fixations (64.9%).

Results and discussion

Fig. 2.4 shows the effect of launch site on the probability that refixations move forward for each word length-landing site pair. Similar to the results of Expt. 1, the GLMM model showed that nearer launch site predicted significantly more forward refixations

($\hat{b} = 0.17$, $SE = 0.03$, $\chi^2_{(1)} = 23.48$, $p < 0.001$, $95\%CI = [0.10, 0.23]$). The same pattern held for both data with landing positions on the left half of the word (3,562 fixations; $\hat{b} = 0.17$, $SE = 0.04$, $\chi^2_{(1)} = 14.04$, $p < 0.001$, $95\%CI = [0.09, 0.25]$), and the right half (271 fixations; estimated from a GLM to ensure convergence) ($\hat{b} = 0.42$, $SE = 0.11$, $\chi^2_{(1)} = 19.25$, $p < 0.001$, $95\%CI = [0.22, 0.66]$). Fig. 2.3 (right panel) shows the GAMM estimation of the effect of launch site with all data of left half aggregated, holding a similar pattern as that of the Dundee corpus. We observe similar results that closer launch sites predicted higher probability of refixations moving forward despite that the PSC differs from the Dundee corpus in several aspects, indicating that the effect of launch site stays robust across different languages and tasks.

2.4. Experiment 3: E-Z Reader

This section aims to show that the standard model does not predict the effect of launch site on direction of refixations. To this end, we carry out the same analyses as the previous section on simulation data from E-Z Reader, a computational model of eye movements in reading that incorporates the standard holistic model of word identification, and always targets refixations to the center of words. In principle, then, all intentional refixations following a fixation on the left half of the word should move forward and those following a fixation on the right half should move backward. Simulations with an implemented version of this model help to ensure that unintentional refixations – saccades intended for another word that happen to become a refixation due to motor error – do not in general change these predictions.

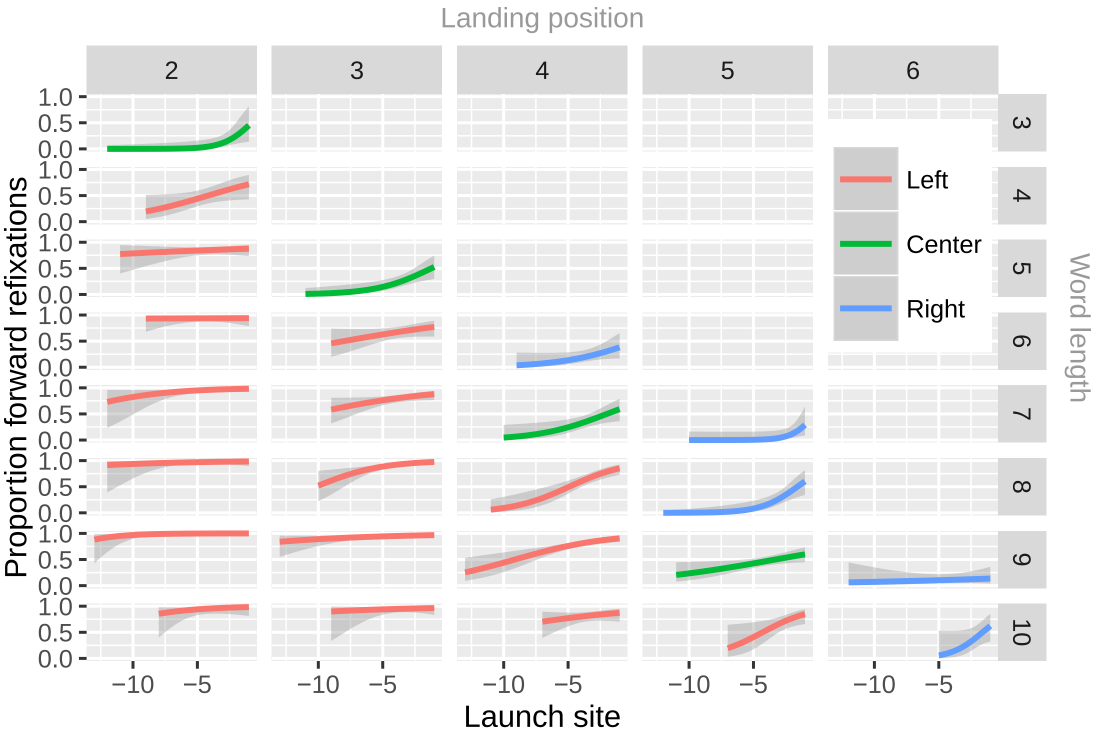

Figure 2.4. Effect of launch site on proportion of forward-moving refixations on data from PSC corpus. Each panel contains data from a combination of word length and landing position, and shows a GAM smoother.

Methods

Data. We used E-Z Reader 10 (Reichle et al., 2009) to generate eye movement data for 100,000 virtual readers reading sentences from the Schilling corpus (Schilling et al., 1998) of single English sentences typical of reading experiments. Each virtual reader was a simulation completed using a Monte Carlo run of the model.

The data cleansing procedure was the same as that in Expt. 1. Out of the 20,189,603 first-pass fixations, 3,417,999 (16.9%) of them were followed by a refixation. Excluding extreme launch sites, fixations landing on initial or final letters of a word, and skipping of the previous word left 1,029,801 fixations. Launch site ranged between $[-15, -1]$.

Statistical analysis. A generalized linear model (GLM) with the same fixed effects as that in Expt. 1 was adopted to analyze the effect of launch sites on refixation direction. Random effects were removed from the GLMM used for Expt. 1 since the virtual readers were simply different Monte Carlo runs with no systematic differences. Excluding word length–landing position pairs where all refixations (or all but 1) moved in the same direction left 899,838 fixations (87.4%).

Results and discussion

Fig. 2.5 shows the effect of launch site on the probability for refixations moving forward. The GLM showed that nearer launch site predicted significantly more forward refixations, $\hat{\beta} = 0.08$, $SE = 0.004$, $\chi^2_1 = 386.66$, $p < 0.001$, $95\%CI = [0.07, 0.09]$. However, this effect was driven by fixations landing on the right half of the word, $\hat{\beta} = 0.10$, $SE = 0.004$, $\chi^2_1 = 542.99$, $p < 0.001$, $95\%CI = [0.09, 0.11]$, while fixations landing on the left half had 99% refixations moving forward and yielded an opposite effect, $\hat{\beta} = -0.33$, $SE = 0.03$, $\chi^2_1 = 147.37$, $p < 0.001$, $95\%CI = [-0.39, -0.27]$. In Fig. 2.6 (left panel) plotted with all data of left half aggregated, the proportion of forward refixations kept as a constant right below 100% and did not vary as launch sites changed.

Therefore, E-Z Reader does not in general predict that closer launch sites should lead to refixations being more likely to go forward, contrary to our observations on the human data, although it can accommodate such a prediction for fixations on the right half of the word. Although this effect on the right half of the word may seem surprising, we note that the predictions we described above for this account only hold for *intentional* refixations. We believe that this effect on refixations on the right half of the word arises from unintentional refixations. Specifically, for a fixation position on the right half of a word, the E-Z Reader model will generally execute one of two behaviors: initiating a saccade to refixate the word or initiating a saccade to move on to the next word. In this case, an intended refixation will target a leftward position (since the center of the word is to the left of fixation) and an intended saccade to the next word will target a rightward position. Which of these two behaviors occurs depends on how quickly the identification (or more technically, L_1) is completed for the current word. Closer launch sites mean that identification of the word will be completed more quickly, which in turn will lead to a greater chance of making a forward saccade intended for the next word. Assuming some of these forward saccades become unintentional forward refixations, this creates exactly the predicted relationship between launch site and refixation direction. For the present purposes, however, the main conclusion here is that the standard model cannot reproduce a general effect of launch site on refixation direction.

2.5. Rational models of reading

In this section, we describe an implemented rational model of refixations, which we will use in the next section to confirm that the intuitively-derived predictions of the rational account for the relationship between launch site and refixations are actually produced by an implemented rational model. Rational models of reading use Bayesian inference to combine visual information with language knowledge (e.g., contextual information). Based on the posterior distribution, eye movements are selected to maximize identification

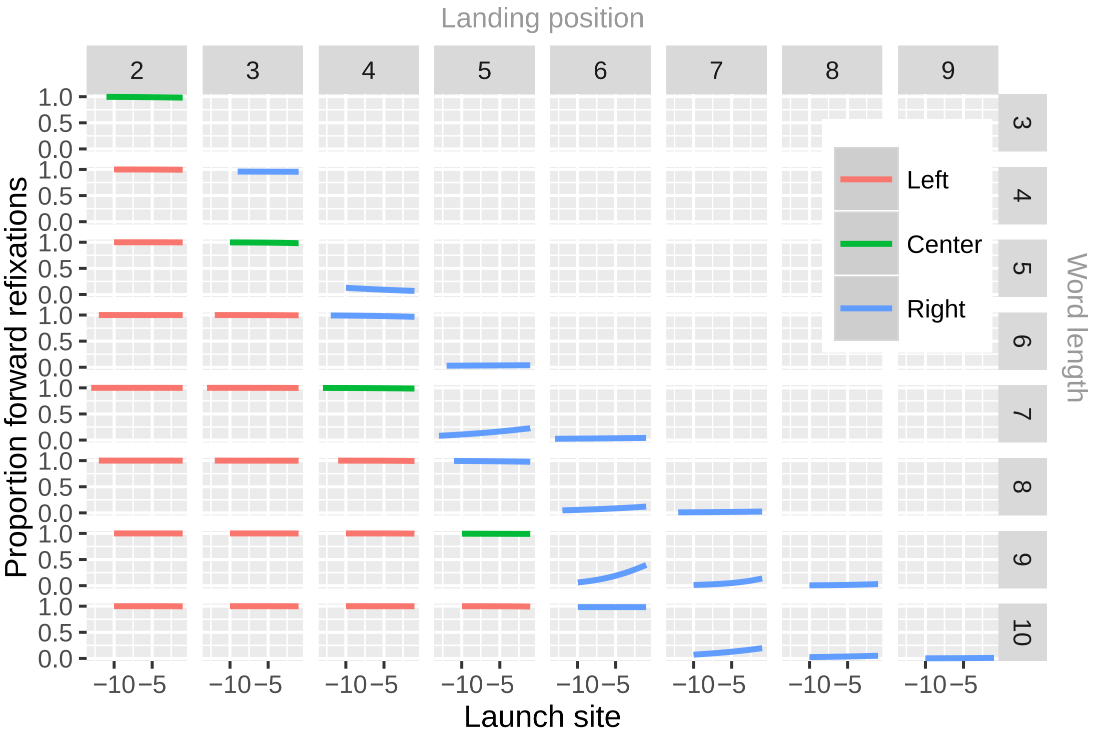

Figure 2.5. Effect of launch site on proportion of forward-moving refixations in data from E-Z Reader simulation. Each panel contains data from a combination of word length and landing position and shows a GAM smoother.

efficiency. The rational model of refixations we describe in this paper also follows this idea, and can be viewed as an application of the more general-purpose rational models of eye movements in reading to the specific situation of refixations. This section introduces the framework of our model.

2.5.1. Word identification as Bayesian inference

Word identification consists of Bayesian inference, in which a prior distribution over possible identities of the text given by its language model is combined with a likelihood term

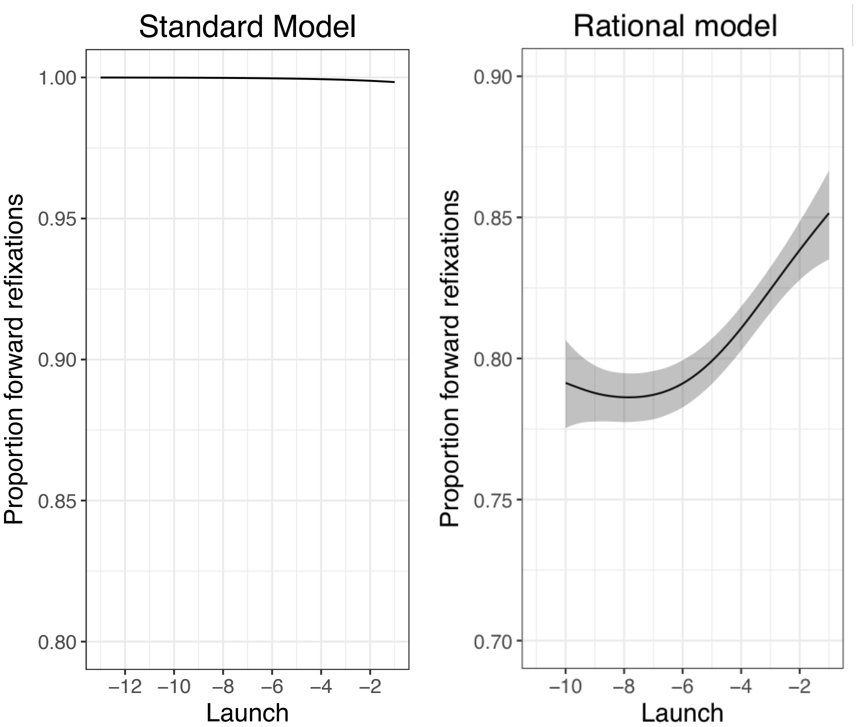

Figure 2.6. GAM estimation of effect of launch site on proportion of forward-moving refixations on data from E-Z Reader (left panel) and rational model (right panel) simulation with initial fixations landing on left-half of the word.

given by 'noisy' visual input at the position of fixation to form a posterior distribution over the identity of the text given all information sources. Formalized with Bayes' theorem,

$$p(w|\mathcal{I}) \propto p(w)p(\mathcal{I}|w) \tag{2.1}$$

where the probability of the true identity of the word being w given uncertain visual input $\mathcal{I}$ is calculated by multiplying the language model prior $p(w)$ with the likelihood $p(\mathcal{I}|w)$ of obtaining this visual input from word w, and normalizing.

In general, the prior $p(w)$ represents reader expectations for words conditioned on the context, but for the present paper, we ignore context and use only a word frequency model for simplicity. The visual likelihood is computed similarly to in (Bicknell & Levy, 2010): each letter is represented as a 26-dimensional vector with a single element being 1 and the rest being 0s. Visual input about each letter is accumulated iteratively over time by sampling from a multivariate Gaussian distribution centered on that letter with a diagonal covariance matrix $\Sigma = \lambda^{-1}I$, where λ is the reader's visual acuity for that letter. Visual acuity depends on the location of the letter in relation to the point of fixation, which is a function of the letter's eccentricity ε. In our model, we assumed that acuity is a symmetric, exponential function of eccentricity:

$$\lambda(\varepsilon) = \int_{\varepsilon-.5}^{\varepsilon+.5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})dx \tag{2.2}$$

with $\sigma = 3.075$, the average of two σ values for the asymmetric visual acuity function ($\sigma_L = 2.41$ for the left visual field, $\sigma_L = 3.74$ for the right visual field) used in (Bicknell & Levy, 2010). In order to scale the quality of visual information, we multiply each acuity λ by the overall visual input quality Λ, which is set to 12 in our simulation (see Expt. 4 below).

2.5.2. Eye movement policy

Based on the posterior distribution on possible identities of the word, eye movement decisions are selected to maximize reading efficiency. For example, the first rational model of reading, Mr. Chips, used this optimizing principle: the model reads input text sequentially, without error, in the minimum number of saccades (Legge et al., 1997, 2002). Specifically, saccades were made to minimize the expected entropy of the current word after the next fixation.

In a more recent rational model of eye movements in reading (Bicknell & Levy, 2010, 2012b), eye movement decisions depend on the uncertainty of the posterior distribution about each letter position. Specifically, given a fixation landing on an unknown character c in position j, the marginal probability m of the most likely character under the posterior is

$$m(j) = \max_{c} p(w_j = c) \tag{2.3}$$

where w_j indicates the character in position j. A high value of $m(j)$ indicates relative confidence about the character's identity, and a low value relative uncertainty. The model then decided between four possible actions based on $m(j)$: continuing to fixate the current landing position, moving backward, moving forward, and ending the reading process.

We use a similar eye movement policy in our refixation model. If the value of the aforementioned statistic $m(j)$ is less than a parameter α, the model chooses to continue fixating the current position. Otherwise, if the value of $m(j)$ is less than the parameter β for some leftward position, the model initiates a saccade to the closest such position. If no such positions exist to the left, then the model initiates a saccade to the closest position to the right for which $m(j) < \alpha$. Once a refixation is executed, the simulation ends. If all $m(j)$ values to the right (left) are above α (β), we decide this word is identified with a satisfactory uncertainty level, and the identification of this word ends. In such a situation, we expect that the eyes move to the next word, which is beyond the current paper's scope of studying refixations.

The actual landing position is the intended fixation position with random motor error: the actual landing position ℓ_i is sampled from a Gaussian centered on the intended target

t_i with standard deviation given by a linear function of the intended saccade distance

$$\ell_i \sim \mathcal{N}(t_i, (\sigma_0 + \sigma_1 |t_i - \ell_{i-1}|)^2) \tag{2.4}$$

for some linear coefficient σ_0 and σ_1.[3] In Expt. 4 in this paper, we follow the SWIFT model in using $\sigma_0 = 0.87$, $\sigma_1 = 0.084$. A refixation occurs if the actual landing site of the next fixation falls on the same word.

2.6. Experiment 4: Rational model

In this section, we analyze simulated data from our rational model of refixations to verify that it does indeed make the prediction that we derived from it intuitively: that refixations would be more likely to move forward for closer launch sites. As described in the previous section, the rational model of refixations we use combines information from previous fixations (including the launch site) to form a posterior distribution on the identity of a word through Bayesian inference. It then makes refixation eye movements to parts of the word about which it is uncertain.

Methods

Model parameters. For the language model component of the word identification model (the prior), we used word frequency information (a unigram model) from the Corpus of Contemporary American English (COCA) (Davies, 2016). For this simulation, we did not optimize the behavior policy parameters to maximize reading efficiency as in (Bicknell & Levy, 2010), but set them manually to what we surmised might be reasonable values

[3] Note that motor error in a rational model has only random error (variance), but not systematic error (bias).

of $\alpha = 0.9$ and $\beta = 0.7$. Future work will optimize them, but we do not expect the qualitative predictions relevant to this analysis to change.

Data. Eye movement data were generated to identify a word. All words were in the most frequent 5,000 words in COCA, and word lengths ranged between $[3, 10]$. Launch site had a range of $[-10, -1]$. For each word length, each possible landing position, and each launch site, 200 trials were run to model the word identification process as when a fixation landed on that landing position, preceded by a fixation on that launch site. In each trial, a word was randomly selected uniformly from words with the same length.

Procedure. Each trial began with a fixation with a duration of 200 time steps on the launch site, in order to represent the visual information obtained prior to fixating the word. Then, the fixation at the landing site began. On each timestep of that fixation, visual information was obtained and integrated with prior information to update the posterior, and then a behavior decision was made: whether to continue fixating, make a refixation, or stop reading (see model description).

Statistical analysis. A GLM with the same fixed effects as that in Expt. 3 was adopted to analyze the effect of launch site on refixation direction. Excluding word length–landing position pairs where all refixations (or all but 1) moved in the same direction left 25,636 fixations.

2.6.1. Results and discussion

Fig. 2.7 shows the effect of launch site on the probability for refixations moving forward. As expected, the GLM showed that nearer launch site predicted significantly more forward refixations, $\hat{\beta} = 0.07$, $SE = 0.005$, $\chi^2_1 = 187.62$, $p < 0.001$, $95\% CI = [0.06, 0.08]$. The

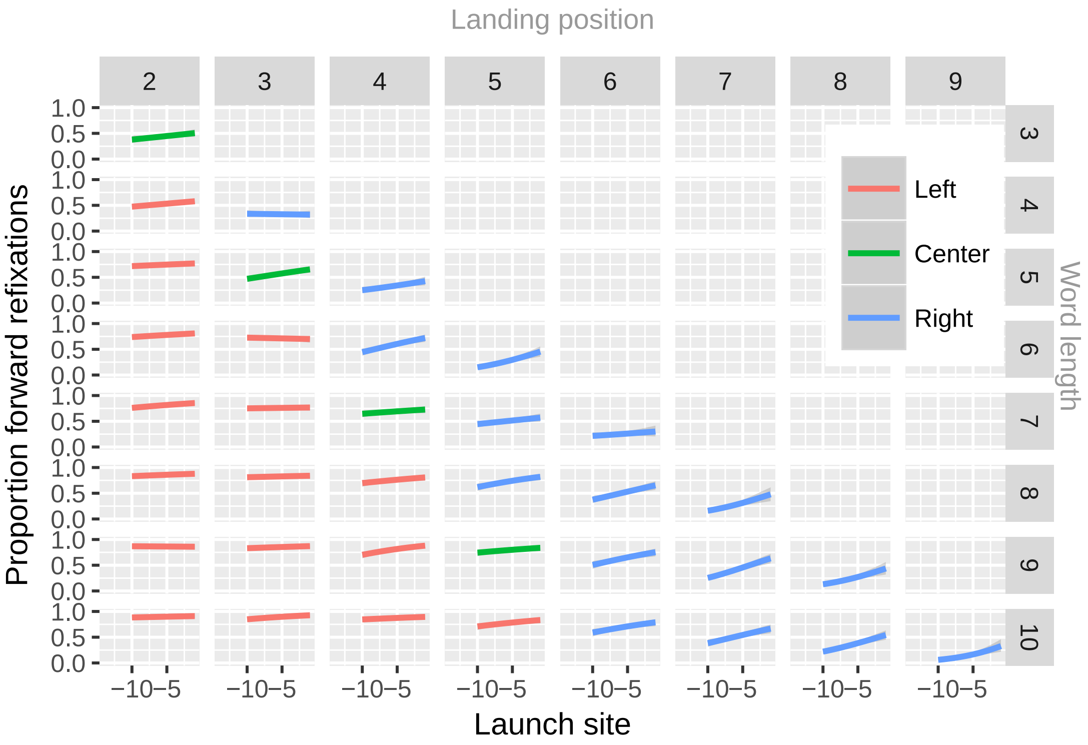

Figure 2.7. Effect of launch site on proportion of forward-moving refixations in data from rational model simulation. Each panel contains data from a combination of word length and landing position and shows a GAM smoother.

same pattern held for both data with landing positions on the left half of the word, $\hat{\beta} = 0.04$, $SE = 0.008$, $\chi^2_1 = 28.85$, $p < 0.001$, $95\% CI = [0.02, 0.06]$, and the right half, $\hat{\beta} = 0.12$, $SE = 0.009$, $\chi^2_1 = 179.38$, $p < 0.001$, $95\% CI = [0.10, 0.14]$. These results confirm that an implemented rational model does indeed make this prediction, which we observed in Expt. 1 and Expt. 2 to hold of human data.

2.7. General discussion

In this paper, we investigated how visual information is used for word identification during natural reading. We compared two accounts: (1) the standard holistic model, in which visual information about the word as a whole is used in word identification, and processing is always most efficient from the center; and (2) a rational model, in which readers combine information from many sources to identify a word constructively, and the fixation location that maximizes identification efficiency depends on what prior information has been obtained. We suggested that these two models make divergent predictions for the possible effects of launch site on where refixations go. Specifically, only the rational model should predict that refixations are more likely to go rightward for closer launch sites. An analysis of a large human eye movement corpus confirmed that this prediction of the rational account holds in human data. Model simulations confirmed that a rational model does indeed predict it, and that at least one of the implementations of the standard model (E-Z Reader) could not accommodate this finding.

These findings seem strongly inconsistent with models in which all intentional refixations target the center of a word, which in turn suggests that the standard holistic model of word identification in reading may be incorrect. However, it is possible to imagine that other refixation targeting schemes could be used even if the holistic model of word identification in reading is correct. For example, even under the standard model, it might be a useful strategy to target a refixation further forward in a word when that word is closer to being identified. Even if there is an efficiency penalty for being away from the center while that word is finished being identified, that penalty might be outweighed by the benefits of being closer to the next word when the reader's attention (soon) turns to it.

While it's possible that such eye movement models could be constructed while maintaining the standard model of word identification, our findings are completely consistent with the predictions of rational models of reading, and suggest that these models should be more fully explored. Here, we focused specifically on how visual information already obtained about a word influences where refixations should go, but rational models predict that the interaction of visual and linguistic information is what should ultimately matter. Future work should test these more complex predictions.

2.8. Acknowledgments

We thank Dr. Reinhold Kliegl for generously sharing Potsdam Sentence Corpus data with us.

CHAPTER 3

A Rational Model of Within-word Eye Movements via Reinforcement Learning

3.1. Introduction

One of the most important channels through which humans interact with the world is reading, the visual processing of language. During reading, readers acquire perceptual input by moving their eyes across the text, identify characters and words by integrating visual input and language knowledge, and extract meaning from this integrated representation. To achieve the goal of comprehending the text, a reader is expected to identify most (if not all) words in the text. We already know from a substantial body of empirical work that readers identify words by utilizing information from various sources, and that readers are able to adopt different eye movement strategies according to the information available. For example, readers take longer to identify low frequency words than high frequency words, whereas they tend to skip words that are short and predictable (see Rayner, 1998 for an extensive review). As word identification is fundamental for visual language processing and is the primary driver of eye movements, understanding how readers decide when and where to move their eyes to best identify a word contributes to understanding reading language comprehension, and also to understanding information processing and decision making in human cognition in general.

By conducting experiments with human participants and recording their eye movements as they read, ample evidence has shown that readers do not move their eyes randomly during word identification. Rather, readers decide when to stop reading and where to move the eyes next depending on the knowledge of the current world (i.e., an eye-mind link exists). The decision of when to stop reading is believed to be associated with cognitive, word-level features, such as word frequency (Kliegl et al., 2004; Rayner et al., 1996; Schilling et al., 1998) and predictability of the word in the context (Balota et al., 1985; Rayner, Slattery, Drieghe, & Liversedge, 2011). The decision of where to fixate next is believed to involve visual features, such as word length (Brysbaert & Vitu, 1998; Vitu, 1991), though this question has been studied to a lesser extent.

In the area of eye movements in reading, researchers usually consider that *when* and *where* to move the eyes involve separate processes (Rayner, 1998). A common assumption is that the time needed to process a word is associated with word difficulty, which can be predicted from word frequency and contextual predictability. Fixating a position other than the word center leads to additional time cost, which is a function of the distance between the fixation position to the word center. To best identify the word, readers always target the word center, although saccade errors add noise to the actual landing position. These assumptions are widely accepted and implemented in eye movement models (e.g. E-Z Reader Reichle et al., 2009 and SWIFT Engbert et al., 2005). The reason behind these modeling assumptions is that visual acuity drops quickly away from the fovea (i.e. a small region of the retina that covers the central two degrees of the visual field) and fixating the word center yields best-quality visual input over the whole word. These models (Reichle et al., 2009; Engbert et al., 2005) involve several parameters, which are estimated by fitting human eye movement corpora, and not surprisingly, eye movement behaviors generated by these models are similar to humans' in terms of several effects (e.g. word frequency effect, word length effect, etc.) and measures (e.g. gaze duration, refixation rate, skipping rate, regression rate, distribution of initial landing position, etc.).

Although the aforementioned models of eye movements fit human data well, they are built upon several assumptions and principles, which merely reflect researchers' expert knowledge about the relationship between eye movement measures and static, word-level features, rather than add insights into how eye movements born out of online cognitive process. A general framework that provides insights of this kind is to consider eye movement control as making decisions to optimize reading efficiency under perceptual, cognitive, and motor constraints. Following the idea of rational analysis (Anderson, 1990), by studying to what extent human eye movements are similar to those generated from an optimized strategy under certain conditions and constraints, researchers can understand what information is available to human readers, and whether human eye movements can be explained as approximation of the optimized strategy.

One study that follows this idea is Reichle and Laurent (2006), which compares human eye movements to an optimal strategy learned from reinforcement learning. Specifically, this model relieves the assumption that readers always target word center, while still maintains other assumptions about word identification, assuming that word difficulty predicts the reading time required for identifying a word. This study examines if human-like patterns of eyes movements can naturally emerge as a virtual reader optimize its behaviors to maximize the total reward of reading a 'sentence' consisting of words of different lengths. Regarding word identification, they consider that the time required to identify a word is a function only of its length and the relative position of the eyes to the word center. This model generates skipping, refixation rate, and first fixation position distributions similar to humans. However, due to the oversimplified word identification model, this model could not yield any linguistic effect.

Another line of research challenges the assumption that a word is identified by being fixated for enough time. They explicitly take visual information into account, and consider that information from various sources is combined to identify a word (Bicknell & Levy, 2010, 2012b; Legge et al., 1997, 2002). Specifically, visual information and language knowledge are combined to yield a probabilistic distribution over the text identity. Successful word identification means that the correct word is the only word with the highest probability, and all other words have probabilities much lower than the correct one. To achieve this, eye movements are performed rationally, obtaining particular pieces of visual evidence that are most useful given the current probability distribution over words. This model of reading explains eye movement behaviors as driven by the rational gathering of visual evidence to best identify the text.

Rational models of reading have been shown to generate predictions that align well with human behaviors for several eye movement phenomena. To name a few, these phenomena include distribution of initial landing position on a word (Legge et al., 1997), word length effect (Legge et al., 1997, 2002), between-word regression (Bicknell & Levy, 2010), word frequency and predictability effect as reflected in several eye movement measures (Bicknell & Levy, 2012b), the effect of launch site on the direction of refixation (Duan & Bicknell, 2017), and word skipping (Duan & Bicknell, 2020). However, there are still places where these models could improve. On one hand, a model that explains both *where* to move the eyes and *when* to move the eyes is missing for identifying a single word, as existing models only focus on *where* to move the eyes (Duan & Bicknell, 2017, 2020). On the other hand, although a model that jointly explains when and where to move the eyes exists, the policy (i.e., the mapping from the agent's current knowledge state to eye movement decisions, where the agent's current knowledge can include information such as the probabilities of possible words and the current fixation location) is constrained (Bicknell & Levy, 2010). Specifically, the policy considers the reader's confidence about the identity of the letter at each position, and moves the eyes to the most uncertain position by referring to two predefined confidence thresholds. Although this policy can approximate optimal decisions by adjusting predefined thresholds, they just cover a restricted policy space, which consists of all policies defined with all possible combinations of confidence thresholds but no any other policies (e.g., say, a policy that requires different confidence thresholds for identifying the current character, any letter to the left, and any letter to the right, and thus needs three confidence thresholds to describe). It is not always clear how the confidence thresholds are determined; they are either determined by referring to an expert's intuition (Bicknell & Levy, 2012b; Duan & Bicknell, 2017), or by comparing the performance of several thresholds (Bicknell & Levy, 2010). Better policies are likely to exist in broader policy space, and only by finding such a policy can we say that this policy is optimal given the reader's knowledge of the current world, without extra constraints.

In this study, we propose a rational model of reading that extends previous rational models of reading to include a policy that maps knowledge of the current word to eye movement decisions by exploring an unrestricted policy space using reinforcement learning (RL). For simplicity, we focus on eye movement behaviors during identifying a single word.

In what follows, we first introduce the eye movement phenomena and explain why existing rational models of reading may not yield full explanations for those phenomena. We then introduce our model by stating the eye movement decision making problem under the framework of RL, and our implementation of this idea using deep reinforcement learning to map states of the current world to actions of eye movements in an end-to-end fashion. We report experiment results using this model to generate eye movements in a word identification task. We evaluate the model's behaviors by comparing them with human data. We provide evidence that the policy learned by RL robustly outperforms other restricted policies and yields eye movement patterns similar to humans' eye movements across different settings.

3.2. Within-word eye movements

As previous studies have provided evidence that eye movement patterns at a whole-word level, such as word length effect and word frequency effect, can be well-explained under a rational model of reading, we attend to fine-grained eye movements at a within-word level. We especially focus on how reading efforts vary by initial fixation position.

3.2.1. Human within-word eye movements

In human reading studies, researchers have found that readers tend to make the first fixation on a word close to the optimal viewing position (OVP), at which position readers' effort needed to recognize a word is minimized (O'Regan, 1992; O'Regan & Lévy-Schoen, 1987). Specifically, researchers have observed that when readers' first fixation lands on the OVP, they identify a word with a shorter period of time and fewer fixations, as reflected in shorter gaze duration (i.e., the sum of all fixations made on a word before making a saccade to another word) and lower refixation rate (i.e., the probability that a word is fixated a second time) respectively, than they start from other positions of the word (McConkie et al., 1989; O'Regan et al., 1984; O'Regan & Lévy-Schoen, 1987; Rayner et al., 1996). This effect is observed across isolated word recognition tasks and natural reading tasks, and is usually presented as a somewhat U-shaped curve with the lowest point located close to and slightly to the left of the word center.

Previous studies also observed complicated interactions between visual information and lexical knowledge regarding the best position to minimize reading times. In fact, the overall pattern that gaze duration is minimized if the reader fixates a letter position close to and to the left of word center does not hold for every word. Rather, OVPs differ for different words depending on the properties of the specific word (Clark & O'Regan, 1999; Farid & Grainger, 1996; Holmes & O'regan, 1992; O'Regan & Lévy-Schoen, 1987). For example, in O'Regan and Lévy-Schoen (1987), researchers conducted an isolated word recognition experiment with words that could be uniquely identified from the first half ("beginning" words) and those could be uniquely identified from the second half ("end" words), and found that they had different OVPs, with the OVP of beginning words shifted leftward and that of end words shifted rightward. Farid and Grainger (1996) found that morphological structure of the stimuli, rather than reading habits or brain lateralization constraints, matters for the location of OVP in an experiment comparing a left-to-right language (French) and a right-to-left language (Arabic). Clark and O'Regan (1999) found that a word could usually be reliably identified if the two letters nearest the location just left of the word center, as well as the very first and the very last letters of the word were known, suggesting that OVPs could be explained from the viewpoint of orthographic constraints. Hyönä, Niemi, and Underwood (1989) manipulated the structure of compound words in Finnish, and found that the eyes initially moved further into a word when the informative information was at the end of the word than at the beginning. These findings are not naturally explained by the account that readers move to the word center for best quality visual input and that readers always target the word center.

3.2.2. Explanations from a rational perspective

A rational model of reading naturally explains why one letter position can be a better position to fixate than other letter positions: given the current knowledge of the probability distribution over all possible word identities as computed from cumulative visual input and linguistic knowledge, the best letter position is expected to yield visual input that best help identify the word. As this word identification model incorporates both visual and linguistic information, we would expect it to predict the following patterns: 1) overall, the effort needed to identify a word (in terms of gaze duration and refixation rate) should be a function of initial landing position and yield an overall OVP close to word center, resulting from visual acuity constraints; 2) overall, the effort needed to identify high frequency words (in terms of gaze duration and refixation rate) should be lower than low frequency words, resulting from linguistic knowledge; and 3) for words with certain properties (we focus on "beginning" words vs. "end" words), the position where gaze duration is minimized should reflect complicated interactions between visual and linguistic knowledge (i.e., should be toward word beginning for "beginning" words and toward word end for "end" words).

Existing rational models of reading face problems in reproducing these effects. The Mr. Chips model considers eye movements to minimize the expected entropy of the word, and is able to reproduce the effect that refixation rate is a function of initial landing position (Legge et al., 2002). However, this model does not have a realistic word identification model, and could not make predictions about reading times. The policy learned through RL in Reichle and Laurent (2006) is able to reproduce the effects that involve visual constraints, but is not expected to predict effects involving linguistic knowledge and especially the interaction between visual and linguistic factors. Existing instances of rational models for eye movements in identifying a single word focus on where the eyes move, and again, do not make predictions about reading times (Duan & Bicknell, 2017, 2020). This only leaves us with one rational model of reading that predicts both reading time and saccade target selection (Bicknell & Levy, 2010, 2012b). This model is expected to predict all the within-word behaviors mentioned above. However, it uses a restricted policy with two hyperparameters, corresponding to the confidence thresholds that readers refer to when they make eye movement decisions, it is not clear if specific choices of the hyperparameters alter the predictions.

To test the rational model's predictions about eye movements at a fine-grained level, and to overcome potential problems of exploring eye movement policies in a restricted space, we implement a rational model of eye movements in reading with a policy learned through RL, and evaluate its behavior by comparing them to humans' within-word eye movements. The idea of learning a policy that maximizes word identification efficiency through RL is inspired by fruitful research that use RL to optimize the strategy to solve motor control problems (Mnih et al., 2015; Sutton & Barto, 2018). In general, RL can solve the problem of deciding which action an agent should take to get the maximum future reward through its interactions with the environment. Therefore, the RL approach only makes an explicit assumption about the goal of reading, and does not introduce extra constraints regarding how actions should be taken at each state. Therefore, RL is appropriate for our goal of finding an optimal eye movement policy in an unrestricted policy space.

3.3. General assumptions and problem statement

3.3.1. Word identification as Bayesian inference

We follow the idea that word identification can be modeled in terms of Bayesian belief updating, in which a prior distribution over possible identities of the text (given by its language model) is combined with a likelihood term (given by 'noisy' visual input at the position of fixation) to form a posterior distribution over identities of the text. Formalized with Bayes' theorem,

$$p(w|\mathcal{I}) \propto p(w)p(\mathcal{I}|w) \tag{3.1}$$

where the probability of the true identity of the word being w given uncertain visual input $\mathcal{I}$ is calculated by multiplying the language model prior $p(w)$ with the likelihood $p(\mathcal{I}|w)$ of obtaining this visual input from word w, and normalizing.

In general, the prior $p(w)$ represents reader expectations for words conditioned on the context, but for the present study, we ignore context and use only a word frequency model for simplicity. The visual likelihood is computed in a way similar to (Bicknell & Levy, 2010): each letter is represented as a 26-dimensional vector with a single element being

1 and the rest being 0s. Visual input about each letter is accumulated iteratively over time by sampling from a multivariate Gaussian distribution centered on that letter with a diagonal covariance matrix $\Sigma = \lambda^{-1}I$, where λ is the reader's visual acuity for that letter. Visual acuity depends on the location of the letter relative to the point of fixation, which is a function of the letter's eccentricity ε. In our model, we assumed that acuity is a symmetric, exponential function of eccentricity:

$$\lambda(\varepsilon) = \int_{\varepsilon-.5}^{\varepsilon+.5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})dx \tag{3.2}$$

with $\sigma = 3.075$, the average of two σ values for the asymmetric visual acuity function ($\sigma_L = 2.41$ for the left visual field, $\sigma_L = 3.74$ for the right visual field) used in Bicknell and Levy (2010). To scale the quality of visual information, we multiply each acuity λ by the overall visual input quality Λ, which is a hyperparameter in our simulation.

3.3.2. The word identification problem

We assume that for a human reader (or an agent) that performs word identification, the task is to move eyes in a way that identifies the word quickly and accurately. In the language of reinforcement learning, a reader agent interacts with an environment in which a word is to be identified. At each time step, the reader agent obtains an observation of the state from the environment, which includes the posterior distribution over words or some form of it, and decides on an eye movement action to take. The environment changes when the reader agent acts on it; for example, the posterior distribution changes after getting a piece of visual input. The reader agent also perceives a reward signal from the environment, indicating how good or bad the current state is. This reward signal encodes both speed and accuracy in some form, as word identification efficiency is usually evaluated on these two aspects. Considering that human readers do not have access to the identity of the true word until they stop reading, the virtual reader should not get accuracy-based reward as well until the end of a trial. The goal of the reader agent is to maximize its cumulative reward by recognizing the word quickly and accurately before stopping reading. The reader agent can use a predefined policy or learn its own policy to achieve this goal.

To make it concrete, we formalize the problem in the way we implement it in this study as below. Note that there can be other implementations as well. Say we have a vocabulary with all words in it having the same length l[1]. We can describe the word identification setup as a Markov decision process, characterized by ($\mathbf{S}$, $\mathbf{A}$, R), where

- $\mathbf{S}$ denotes the set of **states** s. A state is a $27l$-dimension vector, with a l-dimension one-hot vector indicating the current fixation location, concatenated with a $26l$-dimension vector indicating the probability of each character (26 is the number of letters in the English alphabet, and we assume all words only consist of lower-case letters) at each position. Note that this representation is a summary of the posterior in terms of character probabilities at each position. We prefer this representation to the posterior over words, because each word adds to an independent dimension of the posterior, resulting in inefficiency to

[1] We have all words in the same length for two reasons. One is for the ease of modeling, as a letter-based visual representation used in this paper yields state vectors of different dimensions for different word lengths. The other one is that word length has a complex influence on eye movements in reading, and human readers may not have perfect information about word length (Bicknell & Levy, 2012a). Therefore, it is recommended to compare words that all have the same length.

consider relations (e.g. neighbors) among words and increased computational cost as vocabulary size increases.

- **A** denotes the set of **actions** a. Our action space includes $l+2$ actions, consisting of l actions to launch a saccade targeting one of the letter positions in the word, one action to keep fixating the current position, and one action to stop reading and launch a saccade to the first letter of the next word.
- R denotes the reward function. The reward is -1 for every time step as the trial continues, and is a weighted log probability of the true word (denoted as p) when the trial ends (see below for more details about the definition of a "trial"). That is,
$$R(t) = \begin{cases} -1 + w\log(p), & \text{if trial terminates at time } t; \\ -1, & \text{otherwise.} \end{cases}$$
In the end, if a trial ends at time t with the probability of the true word being p, then the return (i.e. sum of the rewards) of this trial is $-t + w\log(p)$.

The interaction between the reader agent and the environment breaks into episodes, which naturally corresponds to trials in a psycholinguistic experiment. Each trial is initialized with a word randomly picked proportional to word frequency in the vocabulary, and an initial landing position randomly sampled from l positions with equal probabilities. The first fixation lands on the initial landing position, initialized with f_1 time steps, during which time visual inputs are gathered from this initial landing position and the posterior distribution of word identities is updated. After that, the agent acts according to its policy. Depending on which action is taken, there are three cases:

(1) If the agent chooses to launch a saccade targeting one of the l letters of the word, then the agent is forced to 1) fixate where the agent actually lands for f_1 steps (if this new position is not outside the word) and 2) fixate the current location for f_2 time steps before moving to the new position, where f_2 corresponds to the time needed for the agent to plan and execute a saccade, known as saccade lag (Rayner, 1998). The agent then lands on this new position, which is drawn from a normal distribution centered on the target position considering random saccade error;

(2) If the agent chooses to stop reading, then a similar process happens, except that the agent launches a saccade that always targets the first letter of the next word;

(3) If the agent chooses to keep fixating the current position, then the agent spends one step on the current location, gets one visual input sample from the current location and updates the posterior.

A trial terminates if i) the agent takes the action to stop reading, ii) time elapsed before taking an action exceeds time limit T (note that a trial can be as long as $T + f_1 + f_2$ because of unintentional refixation due to saccade error), or iii) the actual landing position is outside the word. Since this model considers saccade error and iii) is a condition that could terminate a trial, a trial may terminate intentionally or unintentionally. Specifically, a trial can terminate when the agent targets any of the l positions but actually lands on a position outside of the word, in which case the trial terminates after spending f_2 time steps on the current fixation location. In another case, if the agent chooses to stop reading but the actual landing position is within the word (including letters and the spaces right

Table 3.1. Hyperparameters of the *gym-wordreading* environment.

Notation	Definition	Experiment setting
l	Word length.	7
f_1	Minimum required duration of a fixation before taking any action.	2
f_2	Saccade preparation duration.	1
T	Maximum steps allowed before taking action.	11
w	Weight of log-probability of the true word in the reward function.	3 ('speed') / 10 ('accuracy')

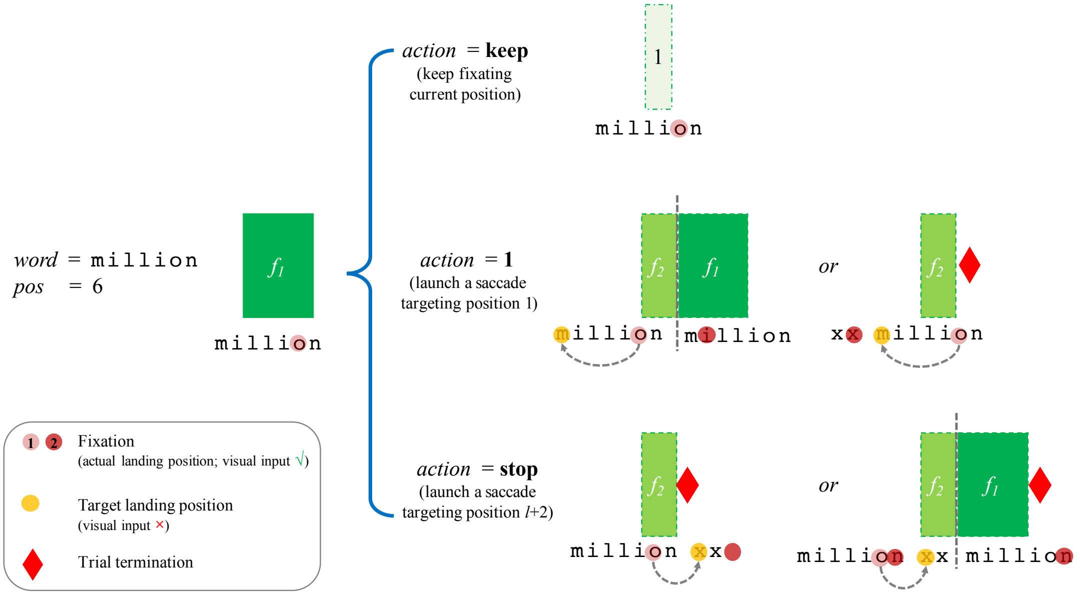

Figure 3.1. An example of the procedure of a trial with true word being 'million' and initial landing position set to 6.

before and after the word), then the agent lands on the actual landing position and gets visual input for f_1 time steps before the trial terminates.

We created an OpenAI Gym (Brockman et al., 2016) interface for this word identification environment, called *gym-wordreading*. Table 3.1 presents the hyperparameters of this environment, and Fig. 3.1 illustrates the procedure of a trial.

3.4. Model implementation

In this section, we first introduce our implementation of a deep reinforcement learning model that performs the word identification task and decides when and where to move eyes. Specifically, the model learns a policy parameterized by a neural network. We then discuss the role of this implementation under the context of other eye movement control models.

3.4.0.1. Policy gradient method. As deciding when and where to move the eyes can be viewed as a motor control problem, we borrow ideas from the field of using reinforcement learning to learn motor control policies (Mnih et al., 2015; Peters & Schaal, 2008). In general, it is straightforward to learn a policy that maps states directly to actions without first computing action values. The idea is to parameterize a policy directly, denoted as π_θ, and the goal is to maximize the expected return $J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} R(\tau)$, where $R(\tau)$ denotes the sum of rewards obtained in a trial with a fixed number of steps, in which the agent acts according to π_θ and generates a trajectory (i.e. a sequence of states and actions) τ. The policy parameters θ can be optimized by gradient descent, such as $\theta_{t+1} = \theta_t + \alpha \, \nabla_\theta J(\pi_\theta)|_{\theta_t}$. The advantages of using a policy gradient are multi-folds: First, the agent can make its policy more greedy over time autonomously, meaning that the policy can start off stochastic to guarantee exploration, and as learning progresses, the policy can naturally converge towards a deterministic greedy policy. Also, with continuous policy parameterization, the action probabilities change smoothly as a function of the learned parameter, avoiding failures due to dramatic change of action probabilities resulting from an arbitrarily small change of estimated action values in pure action-value based methods. There are also times when the policy is simpler than the value function.

Therefore, policy gradient approach is a natural choice for our eye movement control problem.

We utilized an existing Python package, called Spinning Up (Achiam, 2018), to train the reinforcement learning model. Specifically, we used its implementation of Proximal Policy Optimization (PPO). PPO is (roughly speaking) an on-policy, actor-critic method that learns by getting action from a policy (known as the actor) and performing gradient descent based on error signal from an estimated value function (known as the critic), having the advantage of simple to implement and performing at least as well as other policy gradient methods (for more formal details, see Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017). Any other reinforcement learning method may serve our purpose as well as PPO, and we use PPO here just because it is a state-of-the-art reinforcement learning method, it performs well in other motor control tasks, and an implementation of this method is readily available.

3.4.0.2. Model architecture. There are several possible ways of parameterizing the policy into a neural network. A straightforward model architecture is a Multi-Layer Perception (MLP) neural network, in which the input state passes through several fully-connected hidden layers. Alternatively, we can use a convolutional neural network (CNN), which contains convolutional layers and is widely used in image processing. CNN has the advantage of using much fewer parameters than fully connected MLPs, and being able to capture features automatically. In this study, we choose to parameterize our model with a CNN over MLP, because conceptually it is reasonable to represent the visual representation of a word as an image, instead of representing letters at different positions as independent, spatially unrelated features.

The exact architecture, shown schematically in Fig. 3.2, is as follows. We adopt a structure similar to Kim (2014). We first convert the probability of each character at each position into a matrix, each row representing a position by the probabilities of characters as well as the location of current fixation. With l-letter words, this matrix has a shape of $l \times 27$. Then we treat this matrix as an 'image' and perform convolution on it via linear filters. Since rows represent discrete letter positions, we use filters with 'width' equal to the dimension of the position vectors (i.e., 27), and vary the 'height' of the filter, which corresponds to the number of adjacent positions considered jointly. In this study, the convolutional layer convolves 20 filters each of two heights: 2 and 3, with stride 1 and same padding, after which a ReLU activation function is used to induce a feature map. A max-pooling layer follows, where a single number is generated from 2 adjacent positions with stride 2 from each feature map and thus capture the most important feature locally, and these numbers are concatenated to form a feature vector. These features are passed to a fully connected layer with 32 nodes. For the actor, it is followed by a fully connected softmax layer whose output is the probability distribution over actions. For the critic, it is followed by a single node that outputs the estimated value of the given state. This actor-critic object is then trained by PPO to optimize the policy.

3.5. Experiment

3.5.1. Data

Considering that word length influences reading behaviors in a complicated way (Bicknell & Levy, 2012a) and that words of different lengths have different dimensions in a letter-based visual representation we use, we handle words of different lengths separately, and

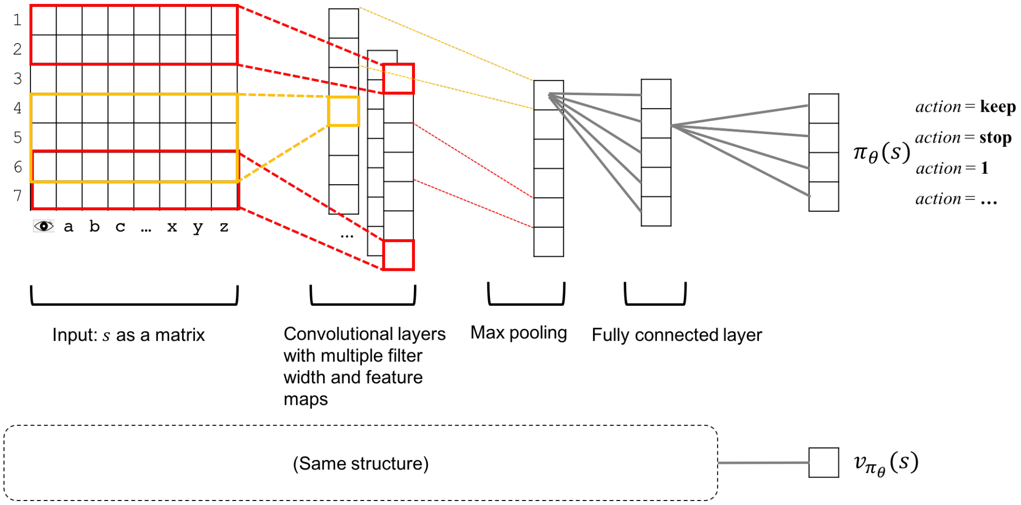

Figure 3.2. Structure of the actor-critic object using a CNN model.

focus on words with a single word length in this study. We included the 3,000 most frequent 7-letter words from Google One Billion Word Benchmark (Chelba et al., 2013) as all possible words in our word identification environment.

3.5.2. Environment settings

In our experiments, we set the environment hyperparameters T to be 11 steps, f_1 to be 2 steps, and f_2 to be 1 step. These numbers are in arbitrary units and do not quantitative fit human reading times, where each fixation usually takes 150 – 300 milliseconds. Rather, they were set to qualitatively demonstrate how such a computational model of reading works. The reason that these numbers were chosen, if any, was to ensure that a word could be identified with a satisfactory accuracy (in our setting, the average probability of true word achieved 80% if word center was fixated through the whole trial) within the given trial length. The trial length was chosen to balance the sensitivity to model difference and the preference for a low computation cost.

For the saccade error function, we assumed the actual landing position to be normally distributed around the target position with a standard deviation given by a linear function of the intended saccade size. We used a saccade error function similar to the Mr. Chips model (Legge et al., 1997), where $pos_{actual} \sim \mathcal{N}(pos_{target}, 0.3|pos_{target} - pos_{launch}|)$.

3.5.3. Training details

We trained agents in this study with 150 epochs. In each epoch, the agent gathers trial experience by acting in the environment with the current policy, and then updates policy gradient once based on the experience. Each epoch allowed a maximum of 25,000 steps to take an action, meaning that the agent learned from at least 2272 trials (if all trials took the maximum possible 11 actions of keeping fixating the current position) before each policy updating. We used the default learning rate of the Spinning Up implementation of PPO. Since we focused on what an RL policy might look like rather than the best possible performance, we did not tune hyperparameters of the settings.

3.5.4. Evaluation

The trained RL agents after the last epoch were evaluated by performing 5,000 trials of word identification in the same environment. We evaluated the policy by comparing the average return of this RL policy against baseline policies, and to what extent this RL policy yielded human-like eye movement behaviors.

3.5.4.1. Baseline policies. As baselines, we implemented the following policies: **random policy (Rand)**, which randomly chooses an action with equal probability; **fixate-center policy (Center)**, which always moves to word center and keeps fixating until one of two conditions is met: 1) the trial reached maximum allowed time steps, or 2) performed the stop-reading action after T' steps, with $1 \leq T' \leq T + f_1 + f_2$ and best T' chosen by grid search in this range; and **alpha-beta policy (AB)** policy, which compares the maximum probability of characters at position j, denoted as $m(j)$, to two hyperparameters α and β: the agent keeps fixating current letter if $m(j) \leq \alpha$; otherwise, the agent first looks leftward and initiates a leftward saccade to the closest position j if $m(j) \leq \beta$, and then looks rightward and initiates a rightward saccade to the closest position j if $m(j) \leq \alpha$; if $m(j)$ exceeds all those thresholds for all letter positions, the agent chooses to stop reading (Bicknell & Levy, 2010). The best combination of α and β was chosen by grid search in $[0.3, 0.5, 0.7, 0.8, 0.9, 0.95] \times [0.3, 0.5, 0.7, 0.8, 0.9, 0.95]$.

3.5.4.2. Within-word eye movements: overall effect of initial landing position. As found in previous research, word recognition time is minimized if readers move their eyes to a position close to and left of the word center, resulting in a U-shaped curve in terms of eye movement measures, especially gaze duration and refixation rate (O'Regan, 1992; Rayner, 1998). To see if the policy learned by the RL agent yielded human-like behaviors, and if the policy took different word properties into consideration, we analyzed three effects.

Firstly, we examined the overall OVP effects in terms of gaze duration and refixation rate. This analysis corresponded to a typical analysis for a word identification experiment, and we expected to find a similar U-shaped gaze duration curve and a similar refixation rate curve as found in human data.

3.5.4.3. Within-word eye movements: word frequency effect. Secondly, we examined if word frequency influenced these curves in a human-like manner. Specifically, we expected that high frequency words took a shorter time to identify than low frequency words, whereas the refixation rate to be not significantly different between them. We also expected that all curves were U-shaped as a function of initial landing position. Although our vocabulary consisted of the top 3,000 seven-letter words and thus all words should be considered as high frequency words, we conducted a median split on log frequency and split words into a high frequency group and a low frequency group. Word frequency was significantly different for these two groups (high frequency: $mean = -4.77$, $SD = 0.46$; low frequency: $mean = -5.70$, $SD = 0.17$; $t = 73.33$, $p < 0.001$).

3.5.4.4. Within-word eye movements: interaction between visual and linguistic factors. Lastly, we examined if RL policy rationally adopted different policies for different words. According to the rational model of reading, readers gather visual information rationally and move their eyes to the most uncertain part of a word. Therefore, a natural prediction is that the position that minimizes reading effort should shift leftward for words with an informative first half of the word compared to words with an informative second half of the word. This pattern was observed in human data (O'Regan, 1992).

One possible way to distinguish words with an informative first half versus words with an informative second half is to see if a word has a unique first half or a unique second half (O'Regan & Lévy-Schoen, 1987). However, a word can have both (e.g., *citizen*, *pyramid*), and how to treat these words is unclear. Other methods that consider word neighbors (e.g. ambiguity analysis in Clark & O'Regan, 1999) require additional assumptions about word identification and thus may involve extra arbitrariness to some extent.

To better capture words that vary in terms of informativeness at different letter positions, we selected words that were more informative in the first half and that were more

Table 3.2. Properties of “beginning” and “end” words.

	“beginning” words	“end” words
Count	150	150
Word frequency (log)	-4.84	-5.05
Average ratio	0.71	1.13
Common prefix/suffix	-ing; -ed	con-; re-
Examples	opening; elected; nursing	control; bathtub; recruit

informative in the second half based on simulation results. Specifically, we consider a letter position informative for identifying a word if the posterior has lower entropy after getting visual samples from this position than from other positions of the word. A word is considered to have a more informative beginning if the average posterior entropy of the first half is lower than the second half, which can be captured by a ratio of these two averaged posterior entropy values. Following this idea, we simulated 30 trials for each word, each letter position, and each time step ranging from 1 to 10, average the posterior entropy for each word and each letter position to reduce random noise, and computed the ratio for each word. We selected words with a ratio lower than 95% of words in the vocabulary as the set of words with an informative first half (“beginning” words) , and words with a ratio higher than 95% of words in the vocabulary as the set of words with an informative second half (“end” words). Properties and examples of these words are shown in Table. 3.2.

3.6. Results

We report results from two simulations with all settings and hyperparameters set the same except for the weight of accuracy in reward function (i.e. w). We set w to be 3 to yield a ‘speed’ setting, in which fast identification is encouraged more than accurate identification, and set w to be 10 to yield an 'accuracy' setting, in which accurate identification is encouraged more than speedy identification. This allows us to see if the RL agent can adjust its policy accordingly, and to see if the policy it learns show robustness across different reward settings.

3.6.1. Reward

First, we ensured that the RL agent learned a stable policy in both a 'speed' setting and an 'accuracy' setting, as illustrated by the temporal evolution of reward during learning in Fig. 3.3.

Fig. 3.4 shows the average reward, average gaze duration, and average probability of the true word of RL policies and other baseline policies on the evaluation trials. We plot the 95% confidence interval of the average reward of all these trials, calculated using bootstrap resampling with 10,000 replicates. As shown in these plots, the RL policy yielded the highest reward resulting from the shortest average gaze duration and the highest average probability of the true word in both the 'speed' setting and the 'accuracy' setting, indicating that an optimal policy exists beyond expert knowledge of eye movements of reading.

3.6.2. Overall effect of initial landing position

Given that the RL agent learned a policy that outperformed other heuristic policies, we further examined what the policy looked like. In this section, we focused on the effect of initial landing position found in human reading experiments, and qualitatively evaluated if the RL policy reproduced these effects.

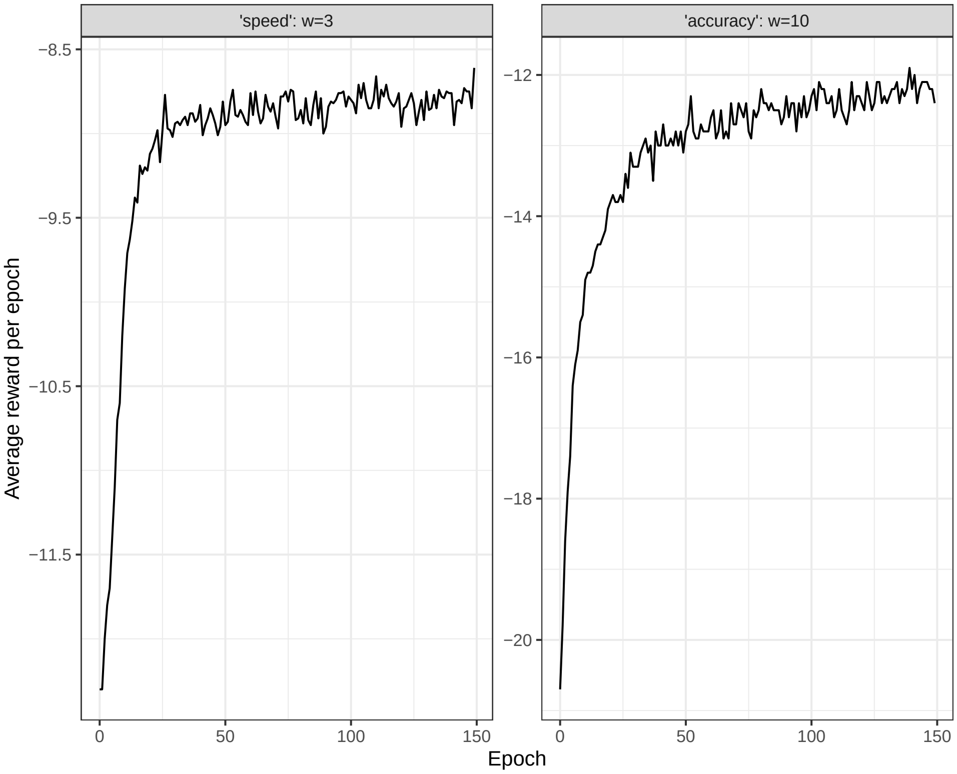

Figure 3.3. Average reward per epoch of training with a CNN model.

As shown in Fig. 3.5, both the gaze duration curve and the refixation rate curve were U-shaped, similar to human reading behaviors observed in human data. Consistent with the intuition that the ‘accuracy’ setting weighed accuracy more than speed, the learned RL policy spent a longer time, and made more refixations than in the ‘speed’ setting.

3.6.3. Word frequency effect

Fig. 3.6 shows gaze durations and refixation rates for high frequency and low frequency words. Similar to human data in Fig. 3.7 (O’Regan, 1992), we observed frequency effects where high frequency words were read faster than low frequency words, and also effects of

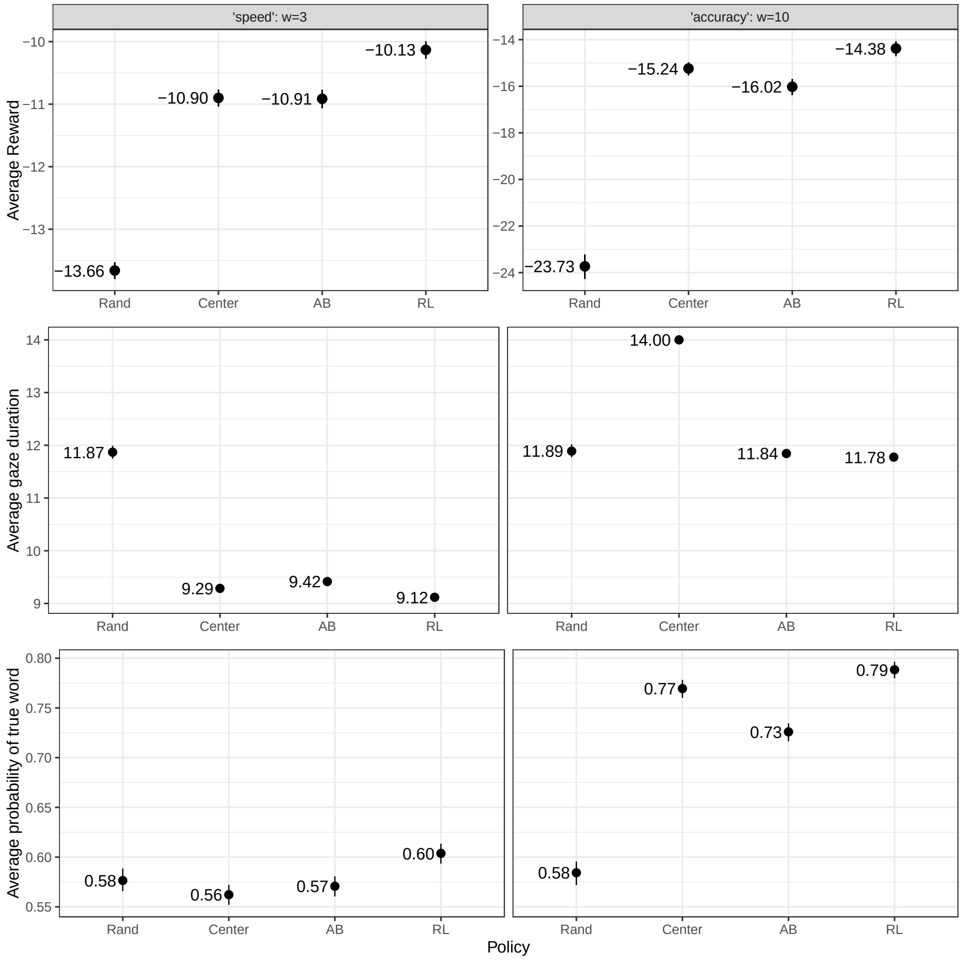

Figure 3.4. Average evaluation reward of different policies in a ‘speed’ setting.

initial landing position where gaze duration curves and refixation curves were U-shaped for both types of words. We did not observe a difference on the refixate rate curve regarding word frequency, which was also the pattern in human reading data. In addition, we observed that for any pair of letters with the same distance to word center, gaze duration and refixation rate were lower for the ones to the left of the word center, consistent with the human reading behavior that the overall OVP is slightly to the left of center, rather than exactly at word center.

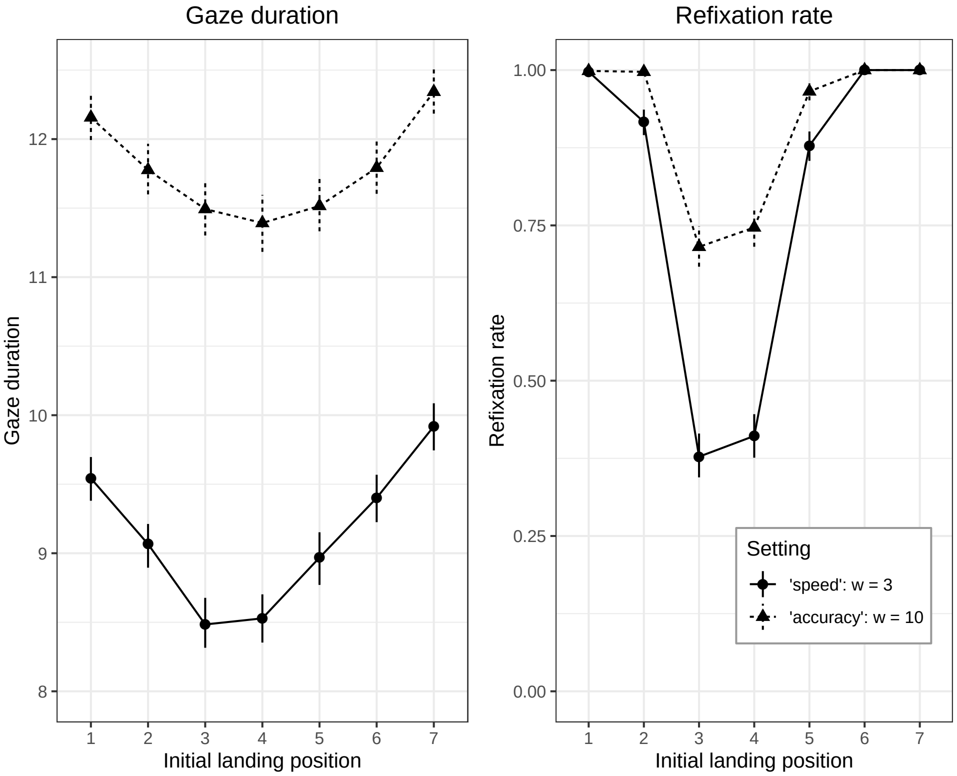

Figure 3.5. Overall gaze duration and refixation rate as functions of initial landing position in ”speed” and ”accuracy” settings.

3.6.4. Interaction between visual and linguistic factors

Fig. 3.8 shows gaze duration and refixation rate for “beginning” words and “end” words. Note that “beginning” words were more frequent than “end” words ($t = 2.78$, $p = 0.006$), and to better align with human data shown in Fig. 3.9, which were collected from a controlled experiment where the frequency of these two groups of words was matched, we removed frequency effects by running regression models with word frequency being the predictor and eye movement measures being the response.

We observed that RL policy produced a human-like pattern, with the optimal viewing position shifted leftward for “beginning” words while shifted rightward for “end” words.

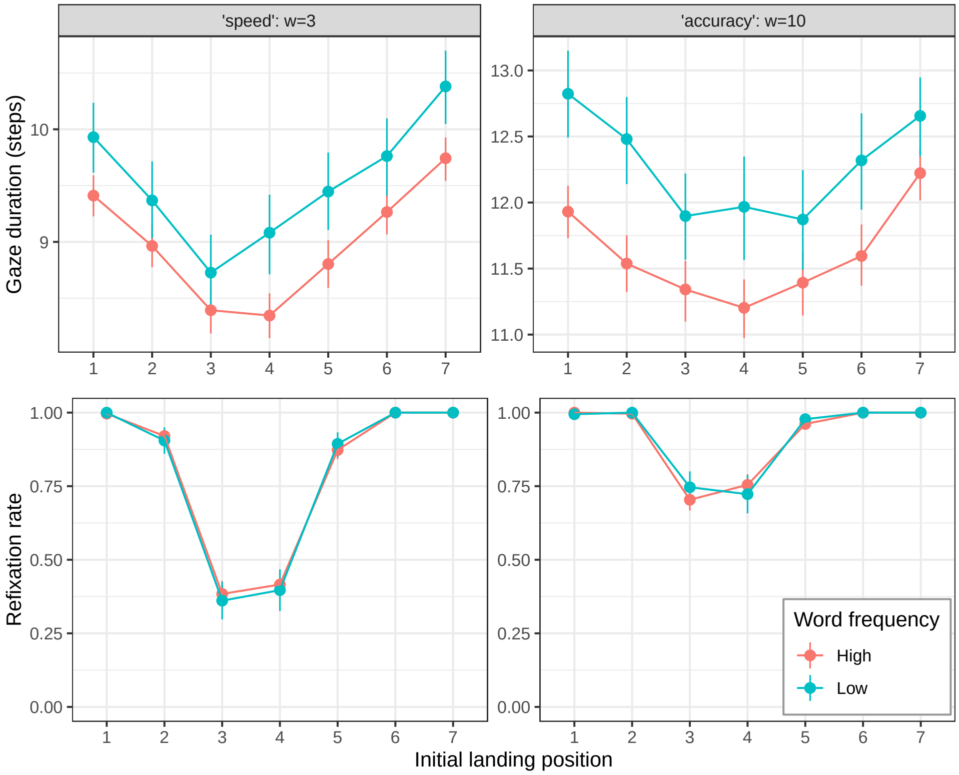

Figure 3.6. Word frequency effect as a function of initial landing position in terms of gaze duration and refixation rate in ”speed” and ”accuracy” settings.

This pattern was most obvious in terms of gaze duration. Although human reading experiments only reported effects on gaze duration, our simulation suggested that refixation rate in the ‘accuracy’ setting may also show an effect. This pattern was interesting because it may suggest that more carefully planned strategy may be used if recognition accuracy was weighed more than speed. Future studies could be conducted to collect eye movement data from human readers and test this prediction.

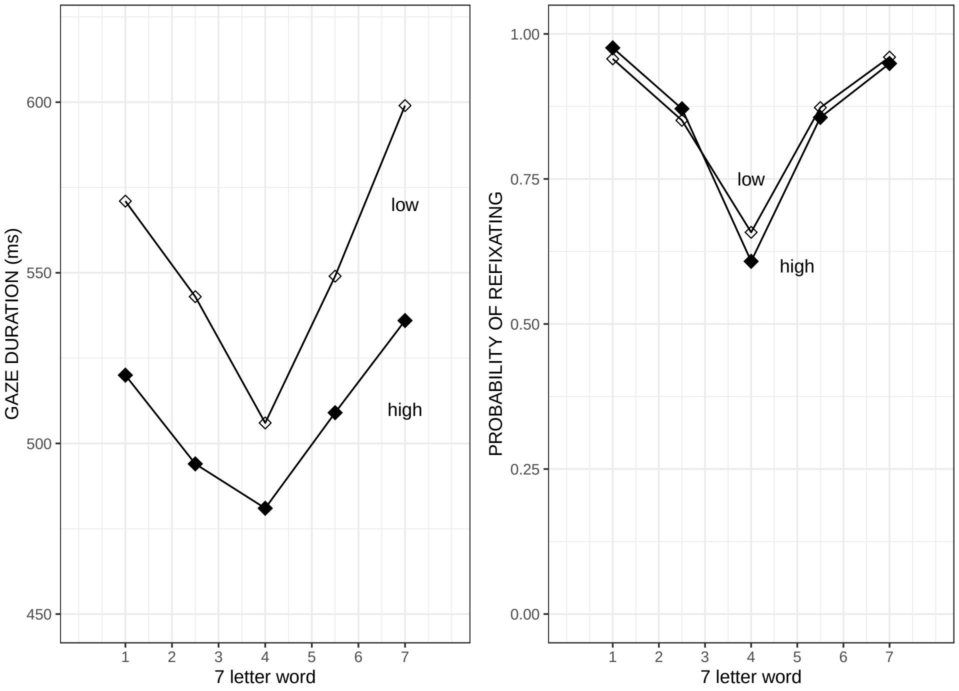

Figure 3.7. Gaze duration and refixation rate as functions of initial landing position and word frequency in human data. Plots are adapted from O'Regan, 1992, Figure 20.2. Note that only fixation positions 1, 2.5, 4, 5.5, and 7 were plotted.

3.7. Discussion

This study presented the first rational model of reading that used deep reinforcement learning to learn a policy that maximized word identification efficiency and tested rational models' predictions at the level of within-word eye movement behaviors. We observed that the RL policy robustly outperformed heuristic policies by achieving higher confidence in a shorter time, indicating that previous heuristic policies were indeed restricted. We also observed that the rational model with an RL policy reproduced human-like behaviors, as evidenced in overall effects of initial landing position, word frequency effects, and in particular different gaze duration patterns for "beginning" and "end" words resulted from

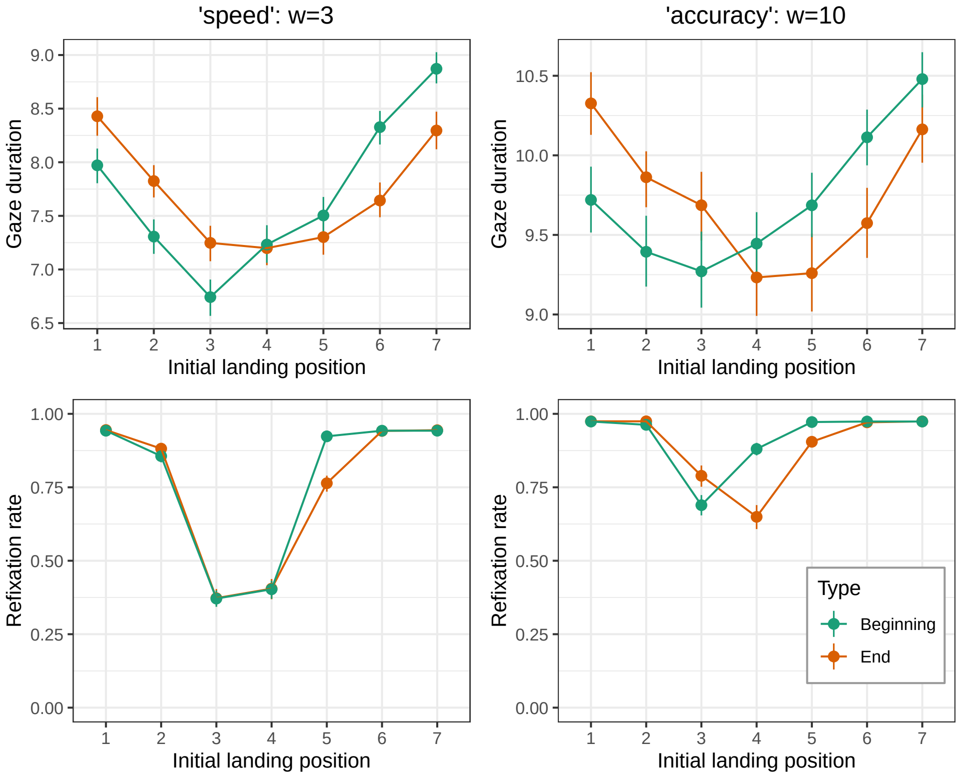

Figure 3.8. Gaze duration and refixation of “beginning” vs. “end” words in ”speed” and ”accuracy” settings.

the integration of visual and linguistic factors, suggesting that rational model of reading were able to explain eye movements at a fine-grained level. These patterns held in both speed and accuracy settings, indicating that our results were invariant to different choices of hyperparameters.

These findings suggest that a rational model of reading provides natural explanations for eye movement decisions regarding when and where to fixate during word identification. Such a framework has advantages in explaining eye movements on two aspects: one is regarding the information based on which eye movement decisions are made, and one is regarding the mechanism of making eye movement decisions. Regarding the information

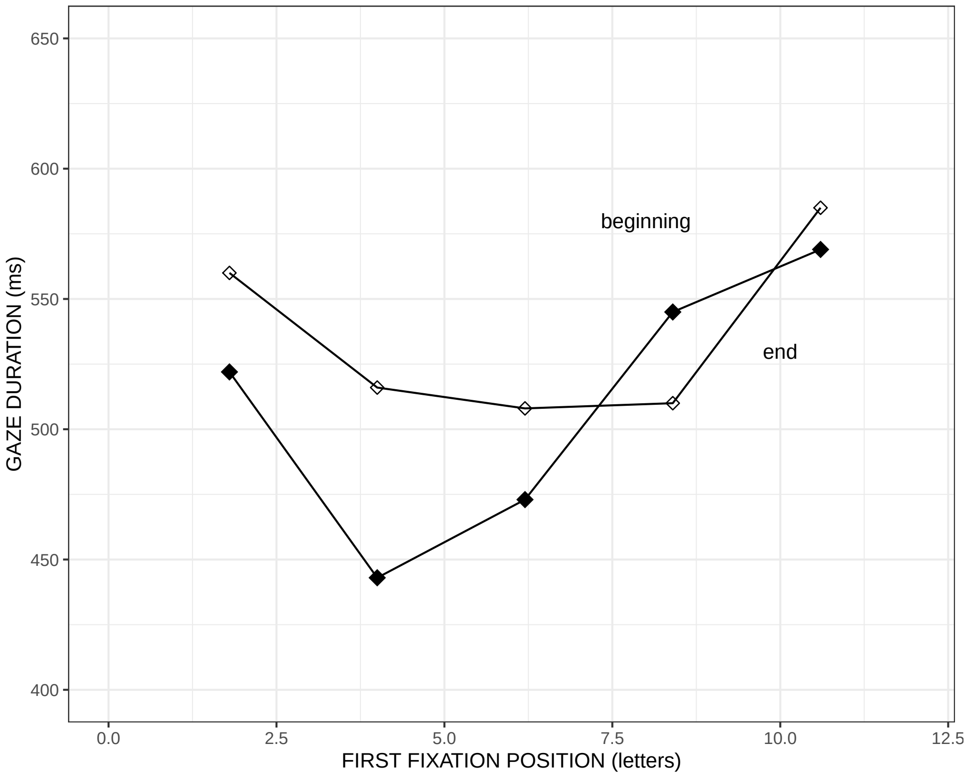

Figure 3.9. Gaze duration for “beginning” and “end” words in human data. Plots are adapted from O’Regan, 1992, Figure 20.3. Note that the data were from a task of identifying words with the length of 10 or 11, and only fixation positions 2, 4, 6, 8, and 11 were plotted.

source, the posterior distribution yielded from combinations of visual and linguistic information turns out to be a useful presentation that leads to reasonable behaviors, suggesting that information from various sources are combined rationally during word identification. This finding echoes previous studies in other eye movement phenomena (Bicknell & Levy, 2010; Duan & Bicknell, 2017, 2020). Regarding the mechanism of eye movement decision making, a policy learned by an RL model that aims to optimize word identification efficiency yields the highest reward and produces human-like eye movement behaviors, suggesting that policies better than fixating word center exist, and humans are likely to adopt similar policies to optimize word identification efficiency.

The RL model learns to decide when to stop reading, showing shorter gaze duration in a ‘speed’ setting than in an ‘accuracy’ setting. It also learns to decide where to fixate and targets positions that yield best word identification performance. Besides its implication that humans make eye movement decision rationally, the RL model itself is also a new application of reinforcement learning that proves to be useful. As shown in this study, deep reinforcement learning can empower the investigation of mechanisms of cognitive processes.

There are also many limitations in this study. Firstly, since human data were collected in an aggregated manner and the OVP effects were coarsely described in previous literature, we did not have the chance to compare more detailed behaviors of the RL policy and human eye movement behaviors. This limited the explanation of a rational model of reading to a relatively coarse level. This could be solved by collecting eye movement data from human participants as they perform a well-designed isolated word identification task, and this would allow more fine-grained comparisons. Secondly, the word reading environment incorporated arbitrary settings, including but not limited to the procedure of eye movements in word identification, hyperparametrs, and saccade errors. It is worth investigating into finding a setting that aligns with human data better (e.g. requiring similar time steps as humans to identify words), such that it becomes possible to look into what components of agent-environment interactions are critical for explaining human behaviors. Lastly, it may not be clear where the current results generalize. Note that we use all words in the vocabulary in both training and evaluation. From the perspective of machine learning, this may implicate over-fitting. However, this is how humans read – the distribution of humans' language knowledge is expected to be exactly the same as what they may encounter, and we do not expect humans' language knowledge to work on language materials they have never seen before (at least at the level of identifying a single word). Regarding using only seven-letter words, we expect all the results of this study to hold in other word lengths as well if different word lengths are treated separately. It would be even more ideal if all word lengths are handled altogether, before which a model of representation that generalizes across word lengths should be built.

This study opens new possibilities for future research. From the perspective of understanding human eye movements, this study provides a framework that directly models eye movement policy from word identification, allowing future research to examine each component of this model in a way similar to examining human cognitive processes but out of the black box. From the perspective of applying deep reinforcement learning to novel situations, this study extends the scope into psycholinguistics, which will potentially benefit from the adoption of reinforcement learning methods, and also serve as a test-bed for their generalization ability.

CHAPTER 4

A Rational Model of Word Skipping in Reading: Ideal Integration of Visual and Linguistic Information[1]

4.1. Introduction

To achieve comprehension in reading, readers move their eyes across the text to obtain the information needed to identify the words. In the past decades, research on eye movements in reading has provided ample evidence that word identification can be seen as the primary driver of eye movements. The reasoning behind this conclusion, however, is based on relatively coarse observations, such as demonstrating that eye movements are sensitive to aggregate variables that are important in word identification (e.g., word length and frequency). Although such a coarse linking hypothesis between word identification and eye movements successfully predicts several reading behaviors, a model of reading that connects eye movements to ongoing language processing in a deeper way could lead to more precise predictions, improved data analysis, and an overall fuller utilization of the eye movement record to advance theories of sentence processing.

One promising model of this type comes from the perspective of rational analysis. The idea is to consider the reading process as one that combines information from various

[1] The contents of this chapter were previously published as Duan, Y., & Bicknell, K. (2020). A rational model of word skipping in reading: ideal integration of visual and linguistic information. *Topics in Cognitive Science*, 12(1), 387-401. doi:10.1111/tops.12485.

sources to identify words and then makes eye movement decisions to maximize identification efficiency (Bicknell & Levy, 2010, 2012b; Legge et al., 1997, 2002). In these rational models of reading, text identification process is modeled using Bayesian inference that combines two sources of information: (1) probabilistic knowledge of the structure of the language, serving as the prior, and (2) uncertain visual evidence, serving as the likelihood. Given a prior and a particular set of visual evidence, probabilistic inference yields a posterior distribution on the text, which specifies the probability of each possible identity of the text. In these models, eye movements are performed to obtain particular pieces of visual evidence. The most efficient, rational reading behavior should use the current posterior distribution on the text identity to determine the most useful time and place to move the eyes next. Therefore, any eye movement behaviors explained by this model of reading can be seen as naturally arising from one source: the rational gathering of visual evidence for text identification.

In contrast, the dominant models of eye movement control in reading tend to use heuristic linking hypothesis between text identification and eye movements (e.g., E-Z Reader, Reichle et al., 2009; and SWIFT, Engbert et al., 2005). For example, in E-Z Reader, eye movements are driven by a word identification process that is represented simply with three discrete states (not identified; partially identified; fully identified). The transitions between these states depend on a certain amount of time having passed, which depends on a few coarse visual and linguistic variables of the word. The timing of the two transitions between the three states depends on a stochastic function of two linguistic variables, the word's frequency in the language and its predictability in context, and one visual variable, the average distance from each of the word's letters to the point of fixation.

Thus, in these models, there should be some effects of each of these three variables on word identification speed. The effects of visual information are that when the eyes are closer to a word, it should be identified faster, and also that if a word is longer, it should be identified more slowly. Similarly, if a word is more frequent in the language or more predictable in context, it should be identified faster.

There are situations, however, where word identification may be affected by more fine-grained information than these three coarse variables. Consider situations where visual information about only the beginning of the word is enough for identification, e.g., seeing the initial letters ‘xyl’ of the word ‘xylophone’ (Hyönä et al., 1989). Similarly, in certain linguistic contexts, a reader only needs to see a few of the initial letters of a word to be confident in its identification, such as in ‘The children went outside to pl…’. Do readers in fact combine more fine-grained information than simply word frequency and word length in the way as predicted by rational models of reading?

As illustrated in the preceding examples, an ideal testbed for these predictions of a rational model is when a word is identifiable with visual information about only part of the word. In natural reading, this situation occurs often in the eye movement behavior of skipping, when a reader moves their eyes past a word without ever having directly fixated it. Intentionally skipping a word is generally modeled as a case in which the reader has identified the word (possibly incorrectly) while still looking at a prior word, and thus makes a saccade that takes the eyes past the word, skipping over it. Since this (implicit) decision about whether to skip the word is made when the reader is fixating a prior word, this is a case when the reader has high quality visual information about only some of the word’s initial letters but does not yet have high quality visual information about the whole word. The amount of visual information the reader has at this time is a function of the *launch site*, the distance from the fixation position to the beginning of the word. In such a situation, both the rational model and the heuristic model predict that how likely a reader is to skip a word should be a function of launch site (amount of visual input), and also of linguistic knowledge (which words are common, and which words are likely in this position). The rational model alone additionally predicts that readers' likelihood of skipping the word will vary depending on the *particular* visual information obtained, and whether that information distinguishes it strongly from its (likely) visual neighbors. Therefore, skipping should be observed to be a complex function of the launch site, the particular word, and linguistic knowledge, in contrast to the heuristic model's predictions of skipping as well-described by coarse visual and linguistic information about the whole word.

Previous empirical research finds that readers' likelihood of skipping a word increases with short word length, close launch sites to the word, high word frequency, and high contextual predictability (Rayner, 1998). Regarding how different sources of information may interact in skipping, studies of skipping short words and especially the word *the* suggest that visual information and word frequency information trump information from the sentence context (Angele & Rayner, 2013; Angele, Laishley, Rayner, & Liversedge, 2014). In particular, Angele and Rayner (2013) manipulate the previews of three-letter verbs being either the verb or the article *the* during sentence reading. They use the gaze-contingent boundary paradigm, in which a preview is replaced by the target verb when the reader's eyes cross an invisible boundary from the left and directly fixate the target. Even though the article *the* is syntactically illegal in the sentence context, the skipping rate is high with this preview, suggesting that word skipping is mainly influenced by parafoveal processing, and that the skipping of *the* is not strongly influenced by the context. Angele et al. (2014) further extends this finding to other three-letter words, with a similar manipulation where the preview mismatches the target word in their syntactic roles in the context. Results show that readers skip the preview of high-frequency words even when they are syntactically illegal in the context, suggesting that the skipping of short words relies on the frequency of the upcoming word more than the fit with the sentence.

Despite these experimental findings suggesting that contextual predictability is not integrated with other information sources, it is still an open question the extent to which readers integrate visual and other types of linguistic constraint (such as word frequency information). Further, this evidence about contextual constraint rests only on these two experiments, which use a relatively small amount and variety of language materials tested and controlled against. The fine-grained predictions of a rational model may be better tested with a set of eye movement decisions that happen in natural reading and that have wide variation in visual and linguistic information available to the reader. The goal of the current paper is to directly test these fine-grained predictions using word skipping, and to gain insights into how readers combine these different sources of information in making skipping decisions.

4.2. Related work

4.2.1. Empirical findings about skipping

At the aggregate level, the effects of visual and linguistic variables on skipping are very robust. Word length is considered to play a more important role than any other factors, as found in a meta-analysis showing that word length explained more variance than word frequency and predictability in regression models predicting skipping rate (Brysbaert & Vitu, 1998). The effect that close launch sites increase skipping rates is also strong and robust (Brysbaert & Vitu, 1998).

As for linguistic variables, there is abundant experimental evidence that skipping rate increases as word frequency increases (Angele et al., 2014; Rayner et al., 1996), and that high predictability leads to high skipping rate (Balota et al., 1985; Rayner et al., 2011). Predictability is usually measured as cloze probability, varying across conditions either with different sentential frames or target words (Balota et al., 1985; Rayner et al., 2011). The effects hold in corpus analysis as well, as Luke and Christianson (2016) find that high target predictability leads to more word skipping for both content and function words. Kliegl et al. (2004) also find significant effect of predictability, word length, and word frequency on skipping rate using regression analyses on Potsdam Sentence Corpus, though they do not include any interactions among these factors.

Several studies have looked into the interactions between visual and linguistic factors at a coarse level. One approach is to analyze linguistic effects on data split by launch sites in post-hoc analysis. For example, Rayner et al. (1996) observe reliable frequency effects on skipping rate at near launch sites (> -5) but not at far launch sites, and White, Rayner, and Liversedge (2005) find a significant interaction between predictability and word length preview overall, which diminishes to a non-effect for far launch sites (near launch sites are defined as those ≥ -3, while far launch sites are those ≤ -4). Another approach to study the interaction of visual and linguistic information is to manipulate parafoveal preview. A preview of the definite article *the* increases readers' skipping rate, even when syntactic constraints do not allow for articles to occur in that position (Angele & Rayner, 2013; Angele et al., 2014). Skipping rates are higher for the preview of a highly predictable word or its visually similar nonword counterpart than the preview of a low-predictability word (Balota et al., 1985), and for the preview of a predictable word than for a visually similar nonword (Drieghe, Rayner, & Pollatsek, 2005). Staub and Goddard (2019) observe that frequency effects on skipping rate are maintained with both valid and invalid previews, but predictability influences skipping only with valid preview. Additionally, English readers only benefit from the preview of a semantically similar neighbor in a highly-constraining context but not in a moderate-constraining context (Schotter, Lee, Reiderman, & Rayner, 2015). More recently, Alhama, Siegelman, Frost, and Armstrong (2019) computed the amount of information available for word identification at different fixation positions by applying a linear filter around the fixation. Using frequent seven-letter words in both English and Hebrew, they find that some words are more readily identified at atypical fixation locations, and readers are sensitive to this information during word recognition.

In sum, previous research has identified visual and linguistic factors that influence skipping by conducting reading experiments and corpus studies. There is also evidence for coarse interactions between aggregate visual and linguistic factors, such as demonstrating that frequency effects only exist if the launch site is close enough to see the word well.

Additionally, these results are also constrained to a small set of well-controlled language materials. A systematic analysis with skipping on a variety of words in a variety of contexts with a variety of launch sites would help gain insights into how visual and linguistic variables interact to identify a word before fixating or skipping it at a fine-grained level.

4.2.2. Other instances of rational models of reading

Previous instances of rational models of reading have provided explanations for several eye movement phenomena. For example, they explain why the initial fixation tends to land near word center and is affected by the launch distance (Legge et al., 2002), why readers often make regressions to previous words (Bicknell & Levy, 2010), and why high-frequency and low-surprisal words yield lower reading difficulty than low-frequency and high-surprisal words (Bicknell & Levy, 2012b). In the field of single word identification, (Duan & Bicknell, 2017) implemented a rational model of refixations, and found that readers rationally make refixations to seek visual information from parts of the word about which the readers are uncertain.

The rational model of skipping presented in this paper has a different focus than previous models. Instead of setting the goal to be identifying a whole sentence, the rational model of skipping here focuses on identifying a single word before directly fixating it. In previous models, the computational cost is high due to recomputing posterior beliefs about an entire sentence after each new piece of visual evidence. The model of skipping is computationally simple, enabling the incorporation of sophisticated models of language knowledge and visual evidence.

4.3. Rational model of skipping

4.3.1. Word identification as Bayesian inference

In our rational model of skipping, word identification uses Bayesian inference, in which a prior distribution over possible identities of the word given by the language model is combined with a likelihood term given by 'noisy' visual input conditional on the fixation position to form a posterior distribution over the identity of the word. Formalized with Bayes' theorem,

$$p(w|\mathcal{I}) \propto p(w)p(\mathcal{I}|w) \tag{4.1}$$

where the probability of the true identity of the word being w given uncertain visual input $\mathcal{I}$ is calculated by multiplying the language model prior $p(w)$ with the likelihood $p(\mathcal{I}|w)$ of obtaining this visual input from word w, and normalizing. Since the shape of the posterior distribution depends on the probability of each word relative to probabilities of other words in the vocabulary, it contains information about how well a word is distinguished from its neighbors.

In general, the prior $p(w)$ represents reader expectations for the next word, and for the present paper, we compare two representations of the prior: a word unigram model (i.e., using word frequency information), which ignores any context information, and a 5-gram model, which conditions on the previous four words of context. The likelihood $p(\mathcal{I}|w)$ represents how likely a piece of visual input is from a word w. For the present paper, we assume that all visual input is obtained only from the final fixation position prior to either fixating the word or skipping it (i.e., the launch site). The visual input obtained about a word consists of independent visual input obtained from each letter in it. Each letter is represented as a one-hot 52-dimensional vector (distinguishing 26 lower- and upper-case letters), with a single element being 1 and the rest being 0. Visual input about each letter is accumulated iteratively over time by sampling from a multivariate Gaussian distribution centered on that letter with a diagonal covariance matrix $\Sigma = \lambda^{-1}I$, where λ is the reader's visual acuity for that letter. Visual acuity depends on the location of the letter in relation to the point of fixation, or eccentricity, which we denote ε. Similar to Bicknell and Levy (2010), we assume that acuity is a symmetric, exponential function of eccentricity:

$$\lambda(\varepsilon) = \int_{\varepsilon-.5}^{\varepsilon+.5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})dx \tag{4.2}$$

with $\sigma = 3.075$, the average of two σ values for the asymmetric visual acuity function ($\sigma_L = 2.41$ for the left visual field, $\sigma_L = 3.74$ for the right visual field) used in Bicknell and Levy (2010). In this paper, we take the scale of σ, the effective width of the visual field, as a free parameter, and experiment with a set of σ scales. In addition, we introduce another free parameter Λ to scale the overall quality of visual information by multiplying it with each acuity λ (see the Experiment section below).

4.3.2. Single word belief updating

Given visual information and linguistic expectations, we may thus compute a posterior distribution over possible identities of the word. Since visual information arrives over time, this is a Bayesian belief updating process, where beliefs are updated as each new piece of visual information arrives. In the single word domain we study here, this Bayesian belief updating process turns out to be relatively computationally simple, and can be implemented as sampling from a multidimensional Gaussian distribution. Say we have a vocabulary of size v, where each word has dimensionality d (here $d = 52\times$ number of characters in the word), and we denote $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_v$ as the vector representations of all the words in the vocabulary. We can represent the current posterior over words at time step t by a $(v-1)$-dimensional log-odds vector $\mathbf{x}^{(t)}$, where each element $\mathbf{x}_i^{(t)}$ represents the log-odds of $\mathbf{y}_i$ relative to the final word $\mathbf{y}_v$. Working with beliefs in this format means that Bayesian inference is just additive in log-odds (no renormalization):

$$\begin{aligned} \mathbf{x}_i^{(t)} &= \log \frac{p(w_i|\mathcal{I}^{(0,\ldots,t)})}{p(w_v|\mathcal{I}^{(0,\ldots,t)})} \\ &= \log \frac{p(\mathcal{I}^{(t)}|w_i)p(w_i|\mathcal{I}^{(0,\ldots,t-1)})}{p(\mathcal{I}^{(t)}|w_v)p(w_v|\mathcal{I}^{(0,\ldots,t-1)})} \\ &= \log \frac{p(\mathcal{I}^{(t)}|w_i)}{p(\mathcal{I}^{(t)}|w_v)} + \log \frac{p(w_i|\mathcal{I}^{(0,\ldots,t-1)})}{p(w_v|\mathcal{I}^{(0,\ldots,t-1)})} \\ &= \Delta\mathbf{x}_i^{(t)} + \mathbf{x}_i^{(t-1)} \end{aligned} \tag{4.3}$$

That is, the log-odds posterior at time step t equals the log-odds posterior at time step $t-1$ (which serves as the prior at time step t) plus the log-odds likelihood. Thus, in an iterative belief-updating context, the log-odds vector begins at a value set by the prior, here the language model, $\mathbf{x}_i^{(0)} = \log p(w_i) - \log p(w_v)$. Then, as each piece of visual information $\mathcal{I}^{(t)}$ arrives, updating beliefs is as simple as adding to $\mathbf{x}^{(t-1)}$ the likelihood log-odds vector for this new piece of information $\Delta\mathbf{x}^{(t)}$, where each element $\Delta\mathbf{x}_i^{(t)}$ gives the likelihood log-odds for that word relative to the final word w_v. For a given true word, vocabulary, and eccentricity, the density function for the likelihood log-odds vector

$\Delta\mathbf{x}^{(t)}$ is a $(v-1)$-dimensional multivariate normal distribution, as each element $\Delta\mathbf{x}_i$ is an affine transformation of $\mathcal{I}$, which is itself a multivariate Gaussian. Specifically, following Norris (2006) and Bicknell and Levy (2010), $\mathcal{I}$ is represented as a vector with dimensionality d, drawn from a multivariate normal distribution with a mean equal to the true word $\mathbf{y}_T$, denoted as $\mathcal{N}(\mathbf{y}_T, \Sigma)$. The co-variance matrix Σ is a diagonal matrix that represents the noisiness of visual input, and the variance of each component is inversely proportional to the processing rate proportion at that letter's eccentricity from fixation. This representation allows the following transformation (Eq 4.4).

$$\begin{aligned}
\Delta\mathbf{x}_i &= \log p(\mathcal{I}|w_i) - \log p(\mathcal{I}|w_v) \\
&= \log p(\mathcal{I}|\mathcal{N}(\mathbf{y}_i, \Sigma)) - \log p(\mathcal{I}|\mathcal{N}(\mathbf{y}_v, \Sigma)) \\
&= [-\frac{1}{2}(\mathcal{I}-\mathbf{y}_i)^T\Sigma^{-1}(\mathcal{I}-\mathbf{y}_i)] - [-\frac{1}{2}(\mathcal{I}-\mathbf{y}_v)^T\Sigma^{-1}(\mathcal{I}-\mathbf{y}_v)] \\
&= \frac{\mathbf{y}_v^T\Sigma^{-1}\mathbf{y}_v - \mathbf{y}_i^T\Sigma^{-1}\mathbf{y}_i}{2} + (\mathbf{y}_i - \mathbf{y}_v)^T\Sigma^{-1}\mathcal{I}
\end{aligned} \tag{4.4}$$

This equation states that $\Delta\mathbf{x}$ can be computed as an affine transformation of $\mathcal{I}$. Since the distribution of $\mathcal{I}$ is multivariate normal, the distribution of $\Delta\mathbf{x}$ is thus also multivariate normal. As a result, the belief-updating process can be implemented as a random walk, where each step is a draw from this multivariate normal distribution.

4.4. Experiment

To test whether readers display signatures of optimal integration across these contexts, we build a computational implementation of an ideal-integration model predicting identification confidence for each skipping decision. We show that these model predictions explain significant variance in human skipping rates when added to a strong baseline model.

4.4.1. Baseline model

4.4.1.1. Data. The English part of the Dundee corpus contains eye movement records from 10 native English-speaking participants as they read through newspaper editorials (see Kennedy & Pynte, 2005, for further details). We included 122,230 observations from the Dundee corpus if they were: 1) a word skipped on first pass (coded as a 1) or a word fixated on first pass (coded as a 0); 2) not adjacent to any blink; and 3) not the first or last fixation on a line. Further, the fixated/skipped word should not 1) contain any non-alphabetical character or be adjacent to punctuation, or 2) follow a word that was skipped or refixated. We excluded observations with far launch sites and long word lengths to ensure enough observations on every level of variations. In the final data, launch sites ranged between [-10, -1], with more than 1000 observations from each launch site, and word length ranged between [1, 8], with the skipping rate being higher than 9% for each word length. The overall skipping rate was 53.9%, resulting from the generally high skipping rate of Dundee corpus, which was over 40% (Demberg & Keller, 2008), and our criterion of requiring the previous word to be fixated, leading to a skipping rate even higher.

4.4.1.2. Model. We analyzed first-pass skipping in the Dundee corpus with a generalized additive mixed-effects regression model (GAMM) predicting skipping from a wide range of variables previously shown to influence skipping, including word length, launch site, word frequency, surprisal, and contextual constraint. We estimated word frequency (log unigram probability) and 5-gram surprisal (log 5-gram probability) with n-gram models (Goodkind & Bicknell, 2018) trained on the Google One Billion Word Benchmark (Chelba et al., 2013), and we measured contextual constraint as the entropy of the 5-gram probability distribution of words in a vocabulary of 20,001 words. We defined the vocabulary to include all words that were in both the Dundee corpus and our language modeling corpus, plus words with frequencies above a cutoff chosen such that the resulting total vocabulary would have about 20,000 words. We also included terms for the previous word's properties such as word length and frequency, and included random intercepts by participants. Crucially, this GAMM allowed for non-linear effects of each of these variables, providing a strong baseline. Table 4.1 shows all the fixed effects in the baseline model.

4.4.2. Rational model

4.4.2.1. Simulation. For each observation in the dataset, we simulated 50 trials using the rational model of skipping for each parametrization of the model. In each trial, a piece of visual information from the launch site is sampled and combined with the linguistic information to generate a posterior distribution of possible identities of the word. As described above, the visual information in this model has two parameters: overall visual input quality Λ and the width of acuity function σ. We used fifteen sets of parameter pairs for the models; these parameters were chosen to be values that spanned a wide part of the parameter space while also respecting the trade-off between width of the acuity function and its overall quality.[2] The linguistic information (prior) in this model is given by either the word frequency (unigram) or 5-gram language models, as used in our baseline model.

4.4.2.2. Analysis. From each trial, we extract the entropy of the posterior distribution (postH) and then calculate the average of postH from the 50 trials for each observation (for each model parametrization). For each parametrization, we add this average postH to our baseline model as a linear predictor. If human readers extract visual and linguistic information in a rational manner, we predict postH to show a significant effect predicting human skipping, even in a strong baseline model, such that skipping is more likely when the posterior entropy is low.

4.5. Results

4.5.1. Baseline model

GAMM results of the baseline model are summarized in Table 4.1. The results confirm previous findings that word length, launch site, frequency, surprisal, and contextual constraint significantly influenced human skipping. Moreover, this baseline model captures non-linear interactions among these predictors, indicating that different sources of information interactively guide skipping at an aggregated level.

4.5.2. Rational model

The partial effects of postH computed from the GAMMs are visualized in Fig. 4.1 (frequency prior) and Fig. 4.2 (5-gram prior), after controlling for all variables in the baseline

[2] If the function is very wide and high quality, the model has too much information about the whole word, whereas if narrow and low quality, the model has almost no information.

Table 4.1. Generalized additive mixed-effects regression model results of the baseline model (note that random slopes for these fixed effects were not included in the model; the model included a random intercept over participants). The GAMM was fitted by REML, and p-values were reported using *summary.gam* function in *mgcv* package (Wood, 2011).

	χ^2	p-value
word length	6026.25	$< 2 \times 10^{-16***}$
launch site	9123.73	$< 2 \times 10^{-16***}$
frequency	527.94	$< 2 \times 10^{-16***}$
surprisal (5-gram)	38.40	$1.01 \times 10^{-6***}$
context entropy	71.16	$8.28 \times 10^{-11***}$
word length × frequency	89.06	$7.73 \times 10^{-16***}$
launch × frequency	36.09	$2.85 \times 10^{-5***}$
launch × surprisal	29.39	$1.13 \times 10^{-4***}$
launch × entropy	66.82	$2.24 \times 10^{-11***}$
word length (word $n-1$)	828.66	$< 2 \times 10^{-16***}$
frequency (word $n-1$)	54.11	$1.62 \times 10^{-9***}$
5-gram (word $n-1$)	127.22	$< 2 \times 10^{-16***}$
context entropy (word $n-1$)	31.68	$5.05 \times 10^{-5***}$
word length × frequency (word $n-1$)	84.69	$1.73 \times 10^{-14***}$

model and additionally including a random slope of postH by participants. The significance of postH when added to the baseline model is reported in Table 4.2. For postH computed from rational models with a frequency prior, the effects are significant in the predicted direction: high postH indicates high uncertainty about the word's identity and is associated with lower skipping rates; these effects are robust to parameter choice and are significant for all parametrizations tested. For postH from rational models with a 5-gram prior, the effects are generally not significant, though they do all trend in the same direction and show the pattern that skipping rates increase as the uncertainty over the word's identity increases, opposite to the predicted direction.

Table 4.2. Significance of averaged entropy of a rational model's posterior distribution when added to the baseline model.

	Prior: Frequency		Prior: 5-gram	
(σ, Λ)	z-value	p-value	z-value	p-value
(1,5)	-2.99	$2.78 \times 10^{-3**}$	1.23	0.22
(1,15)	-2.51	0.012^{*}	1.43	0.15
(1,30)	-2.07	0.039^{*}	2.27	0.024^{*}
(2,5)	-4.49	$7.26 \times 10^{-6***}$	1.15	0.25
(2,15)	-4.22	$2.4 \times 10^{-5***}$	1.67	$0.095^{\cdot}$
(2,30)	-2.75	$6.02 \times 10^{-3**}$	1.96	$0.05^{\cdot}$
(3,5)	-5.76	$8.32 \times 10^{-9***}$	1.23	0.22
(3,15)	-4.92	$8.75 \times 10^{-7***}$	1.56	0.12
(3,30)	-3.88	$1.03 \times 10^{-4***}$	1.04	0.30
(4,5)	-5.98	$2.27 \times 10^{-9***}$	1.16	0.25
(4,15)	-4.22	$2.50 \times 10^{-5***}$	2.15	0.032^{*}
(4,30)	-4.04	$5.36 \times 10^{-5***}$	1.43	0.15
(5,5)	-5.58	$2.37 \times 10^{-8***}$	1.14	0.26
(5,15)	-4.81	$1.55 \times 10^{-6***}$	1.78	$0.076^{\cdot}$
(5,30)	-3.01	$2.65 \times 10^{-3**}$	2.28	0.023^{*}

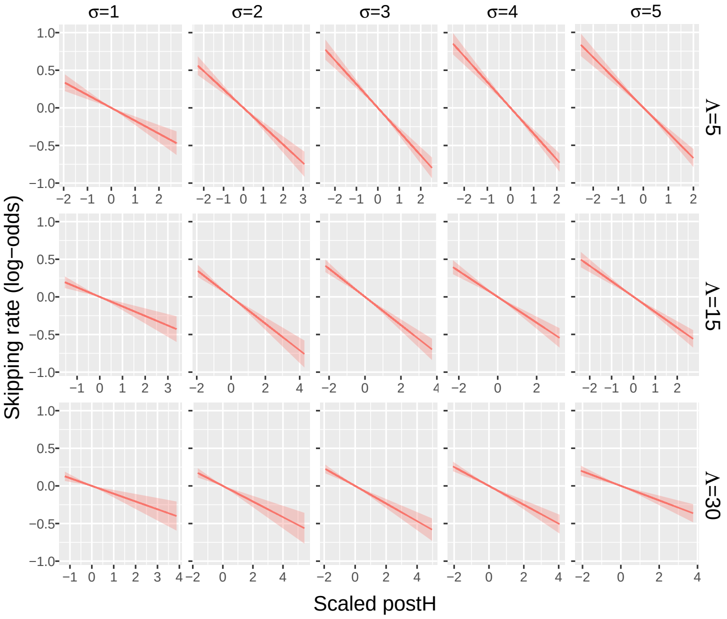

Figure 4.1. Partial effect of postH with a frequency prior in predicting skipping rate.

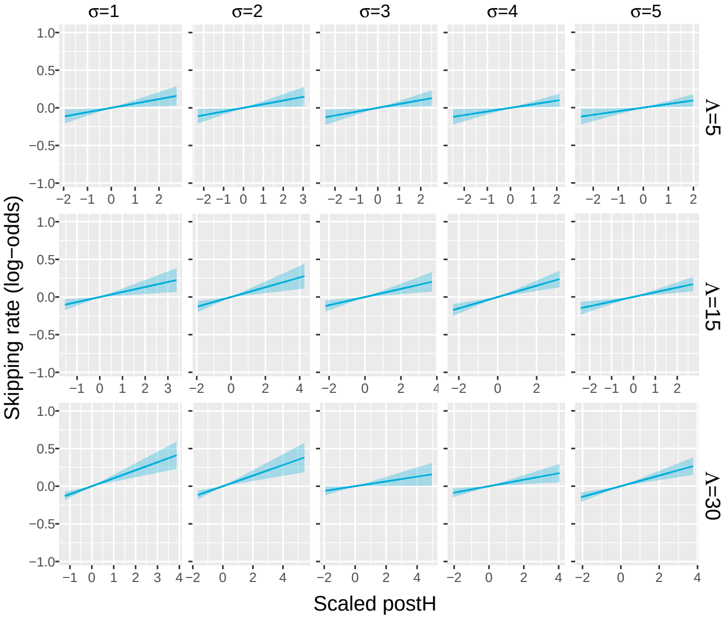

Figure 4.2. Partial effect of postH with a 5-gram prior in predicting skipping rate.

4.6. Discussion

In this paper, we implemented a computational model of skipping that used Bayesian inference to combine visual and linguistic information. We then extracted the entropy of the posterior distribution as a measure of readers' confidence about word identification, and tested whether this measure improved the predictive power of a strong baseline model incorporating aggregate visual and linguistic factors known to influence skipping. Results showed that this postH measure had significant additional effect predicting skipping when extracted from rational models with a frequency prior, but generally not when extracted from rational models with a 5-gram prior. The direction of the effect of postH from models with a frequency prior is consistent with the prediction that low confidence about word identification leads to decreased skipping rate, while the trend of the effect of postH from models with a 5-gram prior is in an opposite direction.

These findings generally provide positive evidence for the rational model's prediction that readers' likelihood of skipping vary depending on the *particular* visual information obtained, and whether that information distinguishes it strongly from its likely visual neighbors according to linguistic knowledge. The predictor, postH, is computed from the posterior distribution of a Bayesian inference model with partial visual information about the word, and therefore captures how likely the word is differentiated from its neighbors in the vocabulary. If the true word is much more likely than its visually-similar neighbors, the postH should be low, while if the true word and its neighbors have similar probabilities, the postH should be high. Such a measure of reader's confidence about word identification is dynamic and hard to capture in factorial experiments, but can be approached through computational simulation. Its significant effect cannot be captured by heuristic models in principle, since postH is assumed to utilize information about how particular words relate to their neighbors regarding the specific visual information obtained about parts of the word.

The observation that postH from a frequency prior better predicts skipping than the 5-gram prior appears to be problematic for a fully rational model of skipping, however: a reader that maximizes usage of all the information available should be better predicted by a model with a prior that conditions on the linguistic context rather than one with an acontextual frequency prior. Instead, this pattern seems consistent with previous findings on the skipping of *the*, which relies on visual and frequency information more than structural information (Angele et al., 2014). This pattern is also consistent with the finding that frequency effects but not predictability effects on skipping survive bad parafoveal visual input, which may be explained by a different time course of frequency and contextual information in making eye movement decisions (Staub & Goddard, 2019).

One alternative possible reason for findings such as ours is that skipping decisions may be made without full knowledge of the context, leading to a poor fit to human data from simulations using a contextual 5-gram prior. Specifically, since saccade programming takes a relatively long time relative to fixation duration and identification/processing of a fixated word continues during this lag, it is plausible that many or most skipping decisions about word n may need to be made before the previous word $n-1$ is fully identified and integrated into the context. In spite of this issue to be further examined, we find that the entropy of a posterior distribution from a frequency prior improves prediction of skipping with average variables, suggesting a complex combination of information sources as predicted by rational models of reading.

4.7. Acknowledgments

This research was supported by a National Science Foundation award 1734217.

CHAPTER 5

Inferring Sentence Comprehension from Eye Movements in Reading

5.1. Introduction

In the past decades, the study of sentence processing has gained fruitful results through the study of humans' eye movements as they read sentences. As has been demonstrated in many studies, there is a tight link between eye movements and cognitive processes in reading, allowing researchers to infer moment-to-moment language processing from eye movements ((Just & Carpenter, 1980; Morrison, 1984; Rayner, 1998; Thibadeau, Just, & Carpenter, 1982), among many others). For example, readers spend more time looking at low-frequency words than high-frequency words, suggesting that eye movements are sensitive to word frequency information (Rayner, 1998). In contrast to the abundant studies regarding the relationship between eye movements and cognitive effort involved in sentence processing, most reading studies remain agnostic about the final outcome of sentence processing: it is almost always assumed that readers recover the correct meaning of the sentence. By 'a same representation' we refer to the correct (or at least, literal) meaning of the sentence. Despite that previous research showed that this assumption is questionable as readers can involve in good-enough comprehension (Ferreira, Bailey, & Ferraro, 2002) and semantic illusion (Sanford & Sturt, 2002), little is known about to what extent is eye movements related to sentence comprehension. In this study, we address this question by predicting comprehension from eye movements using machine learning models and evaluating models' performance in different settings, such that we reveal eye movement patterns that predict sentence comprehension and confirm that eye movements predict comprehension in a way that a rational model of reading expects.

One approach to infer comprehension from eye movements is to use machine learning. In previous studies that have taken this approach, comprehension is measured by readers' performance of answering comprehension questions after reading a few paragraphs (Copeland & Gedeon, 2013; Martínez-Gómez & Aizawa, 2014). A comprehension score can be computed accordingly, allowing predicting non-native readers' language ability from eye movements (Okoso et al., 2015; Yoshimura, Kise, & Kunze, 2015) or text difficulty (González-Garduño & Søgaard, 2018). Regarding how eye movement data are used and which eye movement features are included, there does not seem to be a consensus. In general, these studies use all kinds of eye movement features, such as reading times, number of fixations, number of different types of saccades (e.g. regression to previous text), and even saccade lengths and pupil sizes (Copeland & Gedeon, 2013; González-Garduño & Søgaard, 2018; Martínez-Gómez & Aizawa, 2014; Mishra & Bhattacharyya, 2018). Since these studies usually aim to maximize prediction accuracy rather than to answer how reading process works, the eye movement features they use are not selected based on a particular cognitive theory and are potentially confounded with text properties, preventing these studies from attributing comprehension failures to particular linguistic phenomena or processing mechanisms. In this study, we attempt to address these issues by applying machine learning models to predict comprehension in a well-designed reading experiment, and by examining if integrating eye movement features in a principled way helps machine learning models' predictions.

A theoretical framework regarding the relationship between eye movements and language processing is the rational model of reading. In this model of reading, readers actively acquire visual information by moving their eyes to the most informative part of the text, such that they combine visual information and language knowledge to identify the text efficiently (Bicknell & Levy, 2010, 2012b; Legge et al., 1997, 2002). This process is full of noise, even at the level of word identification that serves as the basis of comprehension. In fact, readers do not always achieve an outcome that is fully consistent with the text; for example, readers tend to adopt higher prior-probability syntactic structures rather than maintain fidelity to text when processing garden-path sentences (Levy, 2011), and they are as easily primed by transposed-letter primes as do identity primes (jugde/judge–JUDGE) (Perea & Lupker, 2003). Rational model of reading can be modeled using Bayesian belief updating, where readers' uncertainty of the identity of the word is measured by the posterior distribution of each possible identity of the text, which is calculated from a prior of probabilistic knowledge of the structure of the language and a likelihood of uncertain visual evidence (Bicknell & Levy, 2010, 2012b). With the assumption that a reader's language knowledge remains unchanged during sentence reading, the posterior distribution largely depends on the visual input. Therefore, if readers obtain enough visual evidence that is critical for word identification, they are more likely to adopt a correct comprehension that is consistent with the text's visual representation rather than one consistent with language knowledge but inconsistent with the text's visual representation.

To examine these predictions of rational model of reading, we focus on a situation in which readers may adopt a comprehension that is largely based on visual information or one that is largely based on language knowledge. This situation is identifying a word with a visually similar neighbor that fits better in the sentence. Two words are neighbors if they have the same number of letters and differ in exactly one letter position (e.g., *glass* and *grass* are neighbors), or their Damerau–Levenshtein edit distance is exactly one, meaning that they require exactly one edit (insertion, deletion, substitution, and the transposition of two adjacent characters) to transform to each other (Slattery, 2009). We say a word has a higher frequency neighbor (HFN) if there exists a neighbor of it with a higher frequency. Slattery (2009) has found processing cost in terms of longer reading times for readers to process words with HFN than words without HFN during sentence reading. This pattern has been explained as that HFN competes with the target word and induces an inhibitory effect. In other words, the uncertainty of the word's identity remains high because both the target word and its HFN receive supporting evidence from information sources (i.e. visual input and language knowledge), and it takes time to acquire visual evidence such that the reader becomes confident that one word identity is more likely than the other.

In this study, we examine to what extent could comprehension be predicted from eye movements and whether eye movements predict comprehension in the way that rational models of reading expect by looking into the identification of words with a HFN during sentence reading. Specifically, we collect eye-tracking data in an experiment in which participants read sentences containing a target word with a HFN, and answer a comprehension question with choices consistent with either the target word or the HFN word. In this case, comprehension is measured in terms of answer accuracy. We then examine a set of *a priori* eye movement features that should be predictive of whether the readers' answers are correct, and check these features' effects using generalized linear mixed effect models. To better evaluate to what extent could eye movement features predict readers' answer accuracy in unseen trials, we implement machine learning models, and evaluate the models' ability to generalize across new participants and regimes. We also examine machine learning models' performance with a feature generated from a rational model of reading, namely the (logit-transformed) probability of the target word, to directly test the rational model's predictions. We conclude the paper with a general discussion.

5.1.1. Related work

5.1.2. Machine learning and eye movements

As eye-tracking techniques become popular, more and more research are conducted in the intersection of machine learning and eye movements in recent years. These research mainly fall into two areas: 1) predicting human readers' comprehension of text from eye movements (Copeland & Gedeon, 2013; Copeland, Gedeon, & Caldwell, 2015; Martínez-Gómez & Aizawa, 2014; Mishra, Dey, & Bhattacharyya, 2017; Mishra & Bhattacharyya, 2018; Okoso et al., 2015; Sanches, Augereau, & Kise, 2017, 2018; Yoshimura et al., 2015), and 2) incorporating eye movement features in state-of-the-art natural language processing (NLP) models to improve these models' performance (Barrett, Bingel, Keller, & Søgaard, 2016; Barrett, Bingel, Hollenstein, Rei, & Søgaard, 2018; Barrett & Søgaard, 2015; Hollenstein & Zhang, 2019). Although these studies focus on different areas rather than the mechanism of eye movements in reading, and implicitly assume that humans' eye movements accurately reflect readers' attention distribution and online cognitive processing, their methods and findings set up valuable reference.

The first set of studies usually collect eye-tracking data from around 10 participants (expect for Copeland et al., 2015, which has 70 participants) reading around 10 pieces of texts (each text usually contains hundreds of words), and measure readers' comprehension or language ability with either a few comprehension questions (Copeland & Gedeon, 2013; Copeland et al., 2015; Martínez-Gómez & Aizawa, 2014), self-rating of readers' subjective comprehension score (Sanches et al., 2017, 2018), standardized language-test scores (Yoshimura et al., 2015), or readers' annotation of the text (Mishra et al., 2017; Mishra & Bhattacharyya, 2018; Okoso et al., 2015). These research use descriptive statistics, correlation metrics, and t-tests to see if eye movement features predict comprehension score or classify readers according to their language abilities. Although eye movements seem to predict readers' comprehension or language ability, comparisons with any baseline model trained with other features (especially text features) are absent.

The second set of studies extract eye movement features from existing eye movement corpora (e.g. Dundee, Kennedy & Pynte, 2005, and ZuCo, Hollenstein et al., 2018) instead of collecting new data. Eye movement features are incorporated NLP models, either by concatenating them with word embeddings (Hollenstein & Zhang, 2019), concatenating them with language model features such as frequency (Barrett et al., 2016), or serving as a constraint of attention (Barrett et al., 2018). In several NLP tasks (e.g. POS tagging, Barrett & Søgaard, 2015; Barrett et al., 2016; named entity recognition, Hollenstein & Zhang, 2019; sentence compression, Klerke, Goldberg, & Søgaard, 2016; and sentiment analysis, Barrett et al., 2018), adding eye movement features improve model performance, though to a relatively weak extent: in Barrett et al. (2018), a baseline model of attention calculated from frequency performs worse than the eye movement attention model by around 1% in terms of F1 score, and in Barrett et al. (2016), the eye movement features alone do not give best performance, and the best feature group (including all features) only outperforms a group of non-gaze features by a 2% in terms of tagging accuracy.

5.1.3. Rational models of reading

Previous instances of rational models of reading have provided explanations for several eye movement phenomena. For example, they explain why the initial fixation tends to land near word center and is affected by the launch distance (Legge et al., 2002), why readers often make regressions to previous words (Bicknell & Levy, 2010), and why high-frequency and low-surprisal words yield lower reading difficulty than low-frequency and high-surprisal words (Bicknell & Levy, 2012b). In the field of single word identification, Duan and Bicknell (2017) implemented a rational model of refixations, and found that readers rationally make refixations to seek visual information from parts of the word about which the readers are uncertain. Duan and Bicknell (2020) implemented a rational model of skipping and found that humans' skipping rate is better predicted by posterior distribution's entropy than by conventional eye movement measures, suggesting that visual and linguistic information is combined rationally.

Compared with previous research, our current study extends the scope of predicting eye movement behaviors itself to predicting language processing outcome from eye movements. It also provides an instance of applying rational model of reading to generate useful features for predictive modeling.

5.2. Experiment and data collection

5.2.1. Participants

Fifty-four students from the University of California, San Diego participated this experiment. They were all native speakers of English with normal or corrected-to-normal vision. They received either course credit or cash as compensation for their time.

5.2.2. Materials and design

Fifty-eight sentences were created, each containing a target word with a high-frequency neighbor (HFN) that was more plausible given the context (Table 5.1). Target words consisted of four to six letters (4 letters: 17 or 29%, 5 letters: 29 or 50%, and 6 letters: 12 or 21%). The target and its HFN were always of the same length, sharing the same first letter, and requiring either a substitution (32 or 55%) or a transposition (26 or 45%) operation to transform to each other. A paired t-test showed that the log word frequency of target words were significantly lower than the log word frequency of HFNs ($p < 0.001$). To examine readers' comprehension of the sentence and especially their comprehension regarding the identification of the target word, each sentence was paired with a multiple choice question. Participants chose only one answer from four choices: one consistent with correct identification of the target word, one consistent with identification of the

Table 5.1. Example sentence. The high-frequency neighbor (HFN) of the Target word (**minuet**) is **minute**.

Sentence:	
I'm really glad that the last **minuet** went by so quickly and I could finally go home.	
Question:	
What was I probably watching to look for cues that I could leave?	
A. an orchestra (correct)	B. a clock (incorrect; HFN consistent)
C. the president (incorrect; unrelated)	D. a weather report (incorrect; unrelated)

target word as its HFN, and two unrelated choices. By measuring readers' comprehension in such an indirect way instead of directly asking readers what the word was with two options, readers were more likely to read in a natural way and were less prone to adopt strategies specific to this task.

5.2.3. Procedure

Participants were calibrated by looking at a random sequence of fixation points presented horizontally across the middle of the computer screen. A fixation cross was presented at the position on the screen where the first character of the sentences would appear. Once a stable fixation was detected within this area, the whole sentence appeared. Participants were instructed to silently read the sentence and to press a key on a keyboard when they finished reading. A comprehension question then appeared, and participants were instructed to choose an answer by pressing a key on the keyboard based on their comprehension of the sentence. Each participant read all 58 sentences, with sentence order randomized for each participant. Participants underwent three practice trials before the experimental sentences.

5.3. Analysis 1: Eye movement features predictive of comprehension

This analysis aims to examine which eye movement features are predictive of comprehension. To this end, we focus on a set of *a priori* eye movement features that are known to be indicative of human language processing, and evaluate these features' effects of predicting comprehension accuracy in generalized linear mixed-effect models.

5.3.1. Method

5.3.1.1. Eye movement features. Based on previous reading research, we selected nine eye movement features, roughly corresponding to two categories: *eye movement features that directly relate to critical visual input acquisition*, as reading times and eye movements that seek visual information (e.g. regression into the target word) increased as text difficulty increased (Copeland & Gedeon, 2013; Martínez-Gómez & Aizawa, 2014; Rayner, Chace, Slattery, & Ashby, 2006); and *eye movement features that indicate general difficulty of processing the whole sentence*, as readers fixated sentence-final words for a longer time and initiated more regressive saccades from there if they detected difficulty in the sentence and carried out reanalysis (Frazier & Rayner, 1982; Von der Malsburg & Vasishth, 2011; Weiss, Kretzschmar, Schlesewsky, Bornkessel-Schlesewsky, & Staub, 2018).

- *Eye movement features that directly relate to critical visual input acquisition.* We include the following features: (1) **Target word total dwell time** is the sum of all fixations' duration on the target word; (2) **Target word skipping** is whether the target word is skipped during first-pass reading or not; (3) **Target word ever fixated** is whether at least one fixation directly lands on the target word or not;

(4) **Regression into target word** is whether the target word is fixated following a regressive saccade or not; (5) **Target character total dwell time** is the sum of all fixations' duration on the target character, which is defined as the character in the target word that needs to be replaced, or the first character that needs to be transposed, to transform to its HFN; (6) **Target character ever fixated** is whether at least one fixation directly lands on the target character or not; (7) **Target character refixated** is whether the target character receives more than one fixation during the whole reading process;

- *Eye movement features that do not directly relate to critical visual input acquisition, but rather indicate general difficulty of processing the whole sentence.* We include the following features: (8) **Total reading time on the sentence's last word** is the sum of all fixations' duration on the last word of the sentence; (9) **Regression out from the sentence's last word** is whether a regressive saccade is launched from the sentence's last word or not.

Besides eye movement features, text features such as word frequency were known to be predictive of comprehension as the more frequent a word was, the more likely a reader was to identify it as that word (Barrett et al., 2016; Martínez-Gómez & Aizawa, 2014). For this reason, we examined the effects of the target word's frequency and its HFN's frequency in predicting comprehension, which is computed from a unigram model trained on the Google One Billion Word Benchmark (Chelba et al., 2013). In addition, as comprehension depended on both text properties and eye movement behaviors (Balota et al., 1985; Rayner et al., 1996), we also examined two more features that focused on the interaction between total reading time on the target word and (i) the target word's frequency or (ii) the HFN's frequency.

5.3.1.2. Data analysis. Estimates were from a generalized linear mixed model (GLMM) for sentence comprehension (1 for correct answer, 0 for incorrect answer), with random intercepts for participants and items using the *glmer* program of the *lme4* package in the R environment for statistical computing. Each predictor was included as a fixed effect in a GLMM separately, yielding 13 GLMMs (9 with eye movement features, 2 with frequency features, and 2 with interactions). We report estimated effect sizes ($\hat{\beta}$ values), standard errors (SE), and z-statistics of fixed effects. Note that our analyses do not serve the goal of statistical inference, meaning that we do not draw conclusion based on whether an estimate is statistically significant or not, and therefore we do not correct for multiple comparisons.

5.3.2. Results & discussion

Trials were excluded from analyses if the same trial was presented for more than once (due to track loss and re-calibration), leaving 3,064 (98%) trials. Data exclusion was evenly distributed across participants and items. Participants incorrectly chose an unrelated choice in 5.0% trials, indicating that they read the sentences carefully; they chose the HFN choice in 19.5% trials, and correctly chose the target choice in 75.5% trials, indicating that readers were likely to misunderstand the target as its HFN. We excluded unrelated choices from further analyses, and only focused on trials in which the target word was either correctly identified or incorrectly identified as its HFN.

GLMM results are summarized in Table 5.2. Although most features showed weak effects, these effects were in expected directions. We observed that 1) fixating the target word at least once increased the probability of correct word identification; 2) high-frequency target words and HFNs were more likely to be correctly identified; and 3) total reading time and word frequency interactively affected word identification, and fixating high-frequency target words for a longer period of time increased the probability of correct identification.

Despite that these *a priori* eye movement features only showed weak effects in predicting comprehension, we confirmed that the effects were in expected directions. These analyses also showed that it was difficult to predict the final outcome of language processing based solely on eye movement features or text features; rather, taking both reading time and word frequency and their interaction into account yielded better predictions.

5.4. Analysis 2: Using machine learning to predict comprehension for new trials

With the knowledge that eye movement features and text features predict comprehension as expected, we now turn to examine to what extent comprehension can be predicted from these features. In this analysis, we frame our task as a classification problem in machine learning, which predicts whether a reader has answered the comprehension question in a trial correctly based on their eye movements during reading. We implement two machine learning models to predict answer accuracy with the *a priori* features we have analyzed in Analysis 1, and compare models' performance across different algorithms, test

Table 5.2. Generalized linear mixed-effects model results of eye movement features and word frequency features.

	$\hat{\beta}$	SE	z-value
Target word total dwell time	0.02	0.06	0.31
Target word skipping	-0.04	0.06	-0.68
Target word ever fixated	0.13	0.06	2.39
Regression into target word	0.06	0.06	1.09
Target character total dwell time	0.02	0.05	0.32
Target character ever fixated	0.05	0.06	0.83
Target character refixated	0.04	0.06	0.69
Total reading time on the sentence's last word	0.09	0.06	1.47
Regression out from the sentence's last word	0.08	0.06	1.30
Target word frequency	0.26	0.21	1.22
HFN frequency	0.49	0.21	2.33
Target word frequency × TD	0.14	0.05	2.77
(Main effect: Target word frequency)	0.25	0.22	1.14
(Main effect: TD)	0.04	0.06	0.71
HFN word frequency × TD	0.09	0.05	1.57
(Main effect: HFN word frequency)	0.49	0.21	2.29
(Main effect: TD)	0.03	0.06	0.50

regimes, and feature sets. This analysis serves two goals: 1) directly examine if comprehension can be predicted from eye movements, and 2) evaluate the generalization ability of machine learning models trained with different features in predicting answer accuracy.

5.4.1. Method

5.4.1.1. Models. We implemented two classification models: a regularized logistic regression (hereafter LR) model and a XGBoost (eXtreme Gradient Boosting, hereafter XGB) model. Both models used eye movement features and text features to predict the binary label that whether a reader correctly comprehend a sentence or not. The LR model used a logistic function to relate the features to the comprehension label. This model was simple, easy to interpret, and less likely to overfit, but it only considered linear effects and tended to underfit. In contrast, the XGB model used gradient boosted trees to make predictions from features. This model was complex and was able to capture nonlinear relations among features, but it was hard to interpret and tended to overfit when dataset was small.

5.4.1.2. Evaluation metric. We measured classifiers' performance using AUC, which referred to the area under the receiver operating characteristic curve (ROC). The AUC measure was equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The AUC ranged between 0 and 1, with a chance level of 0.5, and a higher AUC indicated better performance. This is a standard evaluation metric in machine learning for imbalanced classification problems, such as our dataset. We report the 95% confidence interval of the AUC on the full datset, calculated using bootstrap resampling with 10,000 replicates.

5.4.1.3. Experiment setting. To prevent the machine learning models from overfitting the training data and to increase generalization ability, we tuned hyperparameters of these models by 1) creating folds that consisted of participant (or item) pairs that yielded similar average accuracy and 2) conducting nested leave-two-out cross-validation.

First, to account for the large difference of accuracy among items (ranging from 0.07 to 1) and subjects (ranging from 0.57 to 0.98), we used stratified leave-two-out cross-validation instead of simple leave-one out cross-validation. Specifically, each fold consisted of two items (or subjects), one with a high accuracy score and one with a low accuracy score, such that the average accuracy was roughly the same across different folds. In this way, training folds and test folds preserved similar percentage of samples for each class, increasing the chance that classifiers trained on the training data could generalize to test data as well.

Second, two layers of cross-validation were conducted: 1) each pair of subjects (or items) served as the test set and the rest as training set; and 2) within the training set, we carried out leave-one-pair-out cross-validation to select the best hyperparameters that maximized the evaluation metric (specifically, AUC, see below) on this training set. The advantage of nested cross-validation was that the whole dataset was used in training, validation, and test, which maximized the efficiency of data usage, and was appropriate in our setting with a relatively small dataset.

For both models, we tuned a L2 regularization hyperparameter α with grid search, such that $\min_f \sum_{i=1}^{n}(f(x_i, w) - y_i)^2 + \alpha \left\| w \right\|^2$ served as the loss function, where n denoted the sample size, f denoted the machine learning model (either LR or XGB), and w denoted weights. The regularization hyperparameter ranged between 0 and 30, to the power of 2, and we search with 120 grids. Specifically for XGB, we used a learning rate of 0.03 and a max depth of trees of 6.

5.4.1.4. Test regimes. To examine whether classifiers could generalize to new subjects and new items, we split the data based on subjects and items. Figure 5.1 illustrates the data split for these two test regimes.

5.4.2. Results & discussion

The AUCs of models predicting word identification accuracy with different settings are shown in Fig. 5.2.

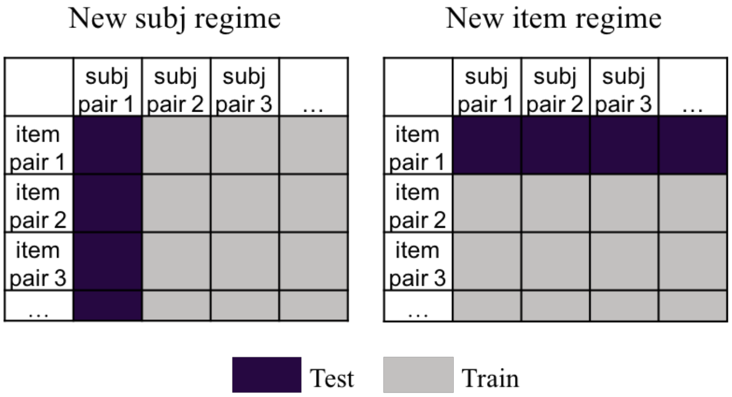

Figure 5.1. Illustration of the data split for new subject and new item regimes.

Regarding the contribution of different features, we found that 1) with eye movement features alone, both LR and XGB were able to predict word identification accuracy better than chance in leave-out-subject regime ($AUC = 0.53$, $95\%CI = [0.50, 0.55]$ with XGB; $AUC = 0.52$, $95\%CI = [0.50, 0.55]$ with LR), and marginally better than chance in leave-out-item regimes ($AUC = 0.52$, $95\%CI = [0.49, 0.55]$ with XGB; $AUC = 0.51$, $95\%CI = [0.48, 0.54]$ with LR); 2) text features predicted word identification accuracy better than chance in almost all settings (except for leave-out-item regime with XGB); and 3) no obvious improvement was observed by adding eye movement features to text features, as indicated by similar performance of text features and all features ($AUC = 0.80$, $95\%CI = [0.78, 0.82]$ vs. $AUC = 0.79$, $95\%CI = [0.77, 0.81]$ in leave-out-subject regime with XGB; $AUC = 0.60$, $95\%CI = [0.57, 0.62]$ vs. $AUC = 0.61$, $95\%CI = [0.58, 0.63]$ in leave-out-subject regime with LR; $AUC = 0.50$, $95\%CI = [0.47, 0.53]$ vs. $AUC = 0.50$, $95\%CI = [0.47, 0.53]$ in leave-out-item regime with XGB; and $AUC = 0.55$, $95\%CI = [0.52, 0.57]$ vs. $AUC = 0.56$, $95\%CI = [0.54, 0.59]$ in leave-out-item regime with LR).

Regarding model performance in different test regimes, we found that it was more difficult to generalize across different items than across different subjects, as indicated by consistently lower performance in leave-out-item regimes than in leave-out subject regimes.

Regarding different machine learning models, we found that XGB yielded the highest performance with text features in the leave-out-subject regime. Roughly speaking, XGB performed better than LR in leave-out-subject regimes, while LR performed better in leave-out-item regimes.

In general, we observed that eye movement features were generally weakly predictive of word identification accuracy, suggesting that eye movements did contain information predictive of comprehension. However, text features were even more predictive, and machine learning models did not benefit from adding eye movement features to text features.

We also noted that predicting word identification accuracy on unseen items were more difficult than on unseen subjects. This pattern suggested that the variation across items were much larger than the variation across subjects in our dataset. Future research could address this issue by reducing the variation between sentences, such as focusing on the comprehension of sentences with high-probability syntactic structures, in which case syntactic structures may be robust to lexical-level variation and invoke similar responses across sentence instances.

5.5. Analysis 3: Rational model simulation

To account for the small size of our dataset and to see if a theory-driven model of combining visual and linguistic information help predict comprehension from eye movements,

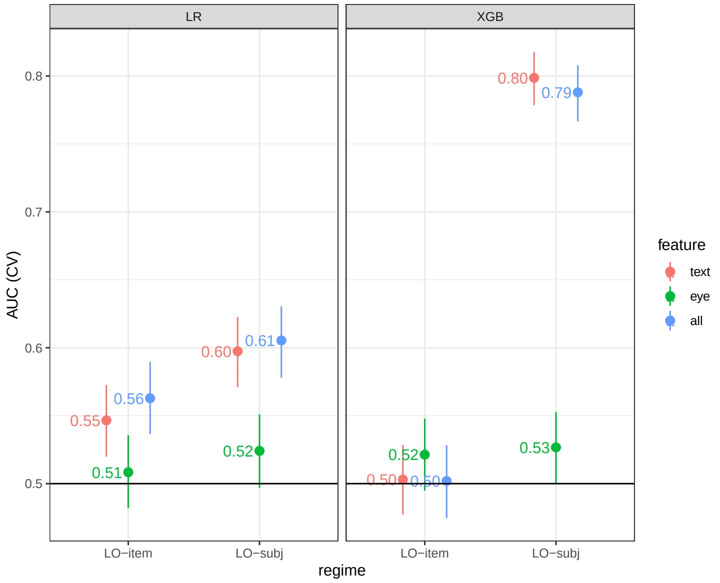

Figure 5.2. AUCs of logistic regression (LR) and XGBoost (XGB) models with text, eye movement, and all features to predict correct/incorrect word identification.

we implemented a rational model to mimic the word identification process, and examined if the probability of the target word predicted word identification accuracy. We repeated Analyses 1 and 2 with this sole predictor. If readers gather visual information rationally during reading, a higher probability of target word in the posterior distribution generated by rational model simulation would predict higher word identification accuracy.

5.5.1. Method

5.5.1.1. Rational model. Following previous rational models (Bicknell & Levy, 2010, 2012b; Duan & Bicknell, 2017, 2020), word identification is modeled as Bayesian inference, in which a prior distribution over possible identities of the text given by its language model is combined with a likelihood term given by ‘noisy’ visual input at the position of fixation to form a posterior distribution over the identity of the text given all information sources.

Formalized with Bayes' theorem,

$$p(w|\mathcal{I}) \propto p(w)p(\mathcal{I}|w) \tag{5.1}$$

where the probability of the true identity of the word being w given uncertain visual input $\mathcal{I}$ is calculated by multiplying the language model prior $p(w)$ with the likelihood $p(\mathcal{I}|w)$ of obtaining this visual input from word w, and normalizing.

In general, the prior $p(w)$ represents reader expectations for words conditioned on the context, but for the present paper, we ignore context and use only a word frequency model for simplicity. The visual likelihood is computed similarly to in (Bicknell & Levy, 2010): each letter is represented as a 52-dimensional vector with a single element being 1 and the rest being 0s. Visual input about each letter is accumulated iteratively over time by sampling from a multivariate Gaussian distribution centered on that letter with a diagonal covariance matrix $\Sigma = \lambda^{-1}I$, where λ is the reader's visual acuity for that letter. Visual acuity depends on the location of the letter in relation to the point of fixation, which is a function of the letter's eccentricity ε. In our model, we assumed that acuity is a symmetric, exponential function of eccentricity:

$$\lambda(\varepsilon) = \int_{\varepsilon-.5}^{\varepsilon+.5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})dx \tag{5.2}$$

with $\sigma = 3.075$, the average of two σ values for the asymmetric visual acuity function ($\sigma_L = 2.41$ for the left visual field, $\sigma_L = 3.74$ for the right visual field) used in (Bicknell & Levy, 2010). In order to scale the quality of visual information, we multiply each acuity λ by the overall visual input quality Λ. In this paper, we adjusted Λ such that the overall average probability of the target word in a given period of time (see below) was close to the overall average accuracy in human response.

For the language model component of the word identification model (the prior), we used word frequency information (a unigram model) trained on the Google One Billion Word Benchmark (Chelba et al., 2013).

5.5.1.2. Simulation. We simulated the process of identifying a target word using the aforementioned rational model of reading. For each item, we used a vocabulary containing only two words: the target word and its HFN. For each trial, we selected fixations that were close to the target word, specifically those landed no more than 5 characters before the word and no more than 5 characters after the word. For each fixation, we simulated a fixation that landed on the same location, and accumulated random visual samples for n time steps, where n equals the duration of this fixation in milliseconds divided by 50 and rounded down to the largest integer no greater than the quotient. After sampling visual evidence with these fixations, a posterior distribution that indicated the probability of the target word and the HFN was generated, and we focused on the probability of the target word. For each trial, such a simulated identification process was repeated for 30 times, and we averaged the logit-transformed probability of the target word to reduce random noise, yielding a single value indicating the rational model's confidence of identifying the target word correctly.

5.5.2. Results & discussion

We first examined the effect of logit-transformed target word probability in predicting identification accuracy using GLMM, following the same procedures as in Analysis 1.

Results showed that high target word probability indicated correct identification as predicted, $\hat{\beta} = 0.084$, $SE = 0.076$, $z = 1.12$.

We then examined if target word probability predicted word identification in unseen trials by carrying out Analysis 2 with logit-transformed target word probability serving as the only feature. The performance was shown in Fig. 5.3. Target word probability predicted word identification accuracy in almost all settings ($AUC = 0.55$, $95\%CI = [0.52, 0.58]$ for leave-out-subject regime with XGB; $AUC = 0.57$, $95\%CI = [0.54, 0.59]$ for leave-out-subject regime with LR; $AUC = 0.53$, $95\%CI = [0.50, 0.55]$ for leave-out-item regime with XGB; and $AUC = 0.56$, $95\%CI = [0.53, 0.58]$ for leave-out-item regime with LR). This pattern suggested that eye movements contained information predictive of word identification accuracy. Note that this one-parameter model performed better than the machine learning models with many more parameters in Analysis 2, suggesting the promise in using principled models for predicting comprehension from eye movements.

5.6. General discussion

The current study examined if comprehension can be predicted from eye movements and if eye movements predict comprehension in the way that a rational model of reading would expect by studying the identification of words with HFNs. Through three analyses, we found that 1) *a priori* eye movement features predicted word identification accuracy in the expected direction, 2) machine learning models trained with just eye movement features predicted significantly better than chance, although the performance also varied depending on different test regimes and did not outperform machine learning trained with text features; and 3) the probability of the target word generated from rational model

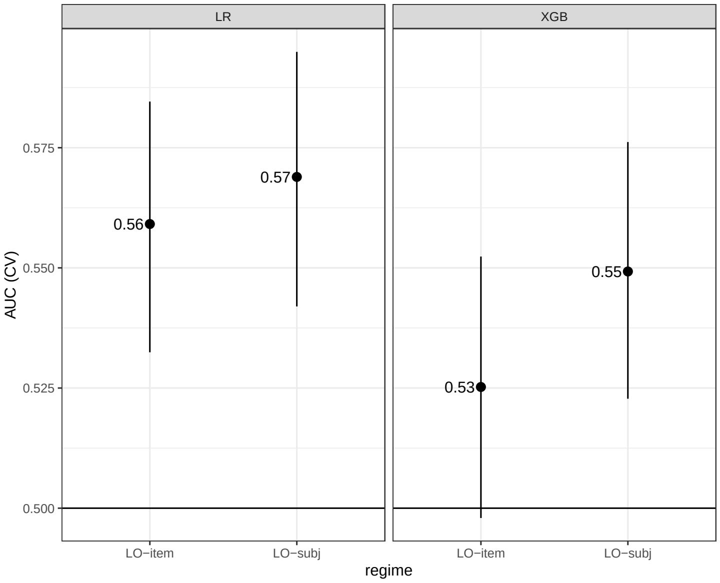

Figure 5.3. AUCs of logistic regression (LR) and XGBoost (XGB) models with rational model's prediction of the target word's probability to predict correct/incorrect word identification.

simulation predicted comprehension accuracy, and machine learning models trained with this feature was able to predict accuracy better than random.

Overall, these results provide supportive evidence that eye movements contain information predictive of comprehension. The current study is the first one, to our knowledge, to examine the relationship between eye movements and comprehension at the level of individual trials. Our findings suggest that studying readers' eye movements helps understanding human language processing, which is taken for granted without careful examination in sentence reading studies. Our finding that generalization across different items is more difficult than generalization across different subjects suggest that researchers should pay close attention to language materials they use in an experiment, as language materials may vary a lot, and may not be representative of the phenomenon that is central to researchers' interest. One possible way to address this issue is to carefully evaluate language materials beforehand, and another way is to use a large set of language materials instead of increasing the number of participants.

The current study also contributes to the research of rational models of reading, especially in the application of using this model to predict human reading processing and outcome. Comparing between our Analysis 2 and 3, the performance of models in Analysis 3 are pretty impressive, as they use just one feature, in contrast to many features in Analysis 2. This result suggests that rational models of reading extract information that is more informative of reading process than traditional eye movement features. In addition, this also provides a promising direction for future analysis of eye movement data, suggesting that it is possible to use a comprehensive metric to measure language processing outcome instead of analyzing multiple eye movement measures, which potentially suffers from multi-comparison issues.

The current study can be improved in many ways. One is the use of machine learning techniques to predict comprehension. Given the large space of machine learning models, in terms of both algorithms and hyperparameters, our choices of models are arbitrary to some extent. For example, instead of evaluating model performance for each regime and feature combination separately, we could optimize overall model performance across different regimes and features. Such change is likely leading to different choices of hyperparameters and exact value of AUCs. However, it is unlikely to influence our main findings that eye movement features alone predict accuracy better than chance and that generalization across items is more difficult than across subjects, as these observations are robust across a simple model (LR) and a highly flexible model (XGB). Instead of the exact numbers of AUCs, we focus more on what these machine learning models reveal regarding the relationship between eye movements and comprehension. It requires collection of a large and representative dataset, delicate feature engineering, skillful algorithm design, and careful model selection to find the best machine learning models that predict comprehension from eye movements to the maximum extent. This is out of the scope of the current study, and the current study should not be interpreted as the best performance that could be achieved by a machine learning model with eye movement features.

In sum, this study confirms that eye movements are revealing of language processing and word identification. We also evaluate the generalization ability of this claim across subjects and items, and point out the importance of taking into account the variation in language materials. Our rational model simulation suggest that comprehension can be better predicted with an integrated metric generated by the Bayesian belief updating model of reading, providing supportive evidence for the perspective of considering eye movements as rational behaviors of gathering visual information for text identification.

CHAPTER 6

Conclusion

6.1. Summary of results

In this dissertation we closely examined the rational account of eye movements in reading and provided evidence that such an account offers qualitatively and quantitatively better explanations for human eye movements than dominant models of eye movements in reading that embedded a standard account of word identification. Chapter 2 showed that humans were more likely to make forward refixations for closer launch sites, which was a pattern predicted by the rational account but strongly inconsistent with the standard account of word identification in reading. These results suggested that visual information obtained from a series of eye fixations was processed constructively to identify a word and to guide eye movements. Chapter 3 demonstrated that readers' within-word eye movement behaviors, in terms of gaze duration and refixation rate, not only depended on initial landing position and word frequency, but also depended on interactions between visual and linguistic knowledge, especially a word's particular visual neighborhood structure. This sensitivity to the structure of the particular word was shown in human data, and was only predicted by a rational model of eye movements. Chapter 4 showed that human skipping decisions were better predicted from the entropy of the posterior distribution of word identification than from a baseline model with coarse features used to model eye movements in dominant models. Finally, Chapter 5 showed that the rational model provided a more robust way to predict readers' comprehension from their eye movements than the way used by dominant models, suggesting that human readers were able to identify words and make eye movement decisions in a rational way.

Taken together, these results suggest that we can understand eye movements in reading as resulting from rationally combining visual and linguistic information to identify the word and using optimized strategy to move the eyes to identify the word quickly and accurately. These findings provide supportive evidence that a rational model of eye movements in reading explains eye movements in word identification, extending previous findings about rational models of reading in explaining sentence-level and word-level eye movement behaviors to a more fine-grained level. Moreover, we propose a new policy learned through deep reinforcement learning that maps readers' real-time state of knowledge about the world to eye movement decisions and show that this policy outperforms existing heuristic policies, providing a useful tool and framework for future research of eye movements in reading.

6.2. Future directions

A potential future direction is to further explore how visual and linguistic information interact in word identification by using different visual parameters, or even using different visual representations. We did not tune visual parameters to reflect human's vision; we reported qualitative similarities between human eye movements and our models' behaviors in two of the four studies. Fitting those parameters to reflect human visual processing rate would help evaluate the models' behaviors quantitatively. In addition, our visual representation was simple and did not take visual similarities among letters into consideration, which can be improved to be more realistic by incorporating a character confusion matrix, or even using an image-based representation for each letter.

Another direction is to go beyond linguistic priors generated from language models and incorporate other sources of linguistic information, such as syntactic information and semantic information. These knowledge have been proved to alter humans' expectations about upcoming words in real-time language processing (DeLong, Urbach, & Kutas, 2005; Kuperberg & Jaeger, 2016), which corresponds to the idea of the prior in the Bayesian model of word identification. Our current n-gram representation of the prior is not expected to capture eye movements observed in psycholinguistic experiments with manipulation of syntactic or semantic cues. Extending the model of eye movements in reading in this direction could yield models that help evaluate whether the rational model can still explain eye movements in reading under the influence of higher-level language processing.

A third direction is to thoroughly evaluate the optimal eye movement policy learned from reinforcement learning. Besides the patterns of within-words eye movements in terms of gaze duration and refixation rate, we can further examine how the policies differ depending on specific knowledge of the world, and thus shed light on potential structures of the knowledge used in eye movements decision making. In general, this new tool of reinforcement learning can be used in various ways, and help understand eye movement decision making both in reading and also in other situations involving eye movements.

References

Achiam, J. (2018). Spinning Up in Deep Reinforcement Learning.

Alhama, R. G., Siegelman, N., Frost, R., & Armstrong, B. C. (2019). The role of information in visual word recognition: A perceptually-constrained connectionist account. In *The 41st annual meeting of the cognitive science society (cogsci 2019).*

Anderson, J. R. (1990). *The adaptive character of thought.* Psychology Press.

Angele, B., Laishley, A. E., Rayner, K., & Liversedge, S. P. (2014). The effect of high-and low-frequency previews and sentential fit on word skipping during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1181–1203. doi: 10.1037/a0036396

Angele, B., & Rayner, K. (2013). Processing the in the parafovea: Are articles skipped automatically? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 649–622. doi: 10.1037/a0029294

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, *17*(3), 364–390. doi: 10.1016/0010-0285(85)90013-1

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., & Søgaard, A. (2018). Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 302–312).

Barrett, M., Bingel, J., Keller, F., & Søgaard, A. (2016). Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 579–584).

Barrett, M., & Søgaard, A. (2015). Reading behavior predicts syntactic categories. In *Proceedings of the nineteenth conference on computational natural language learning* (pp. 345–349).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1168–1178). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=1858681.1858800`

Bicknell, K., & Levy, R. (2012a). Why long words take longer to read: the role of uncertainty about word length. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics (cmcl 2012)* (pp. 21–30).

Bicknell, K., & Levy, R. (2012b). Word predictability and frequency effects in a rational model of reading. In *Proceedings of the 34th annual conference of the Cognitive Science Society* (pp. 126–131).

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). *Openai gym.*

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In *Eye guidance in reading and scene perception* (pp. 125–147). Elsevier. doi: 10.1016/B978-008043361-5/50007-9

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Clark, J. J., & O'Regan, J. K. (1999). Word ambiguity and the optimal viewing position in reading. *Vision Research*, *39*(4), 843–857.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, *49*(2), 602–615.

Copeland, L., & Gedeon, T. (2013). Measuring reading comprehension using eye movements. In *Cognitive infocommunications (coginfocom), 2013 ieee 4th international conference on* (pp. 791–796).

Copeland, L., Gedeon, T., & Caldwell, S. (2015). Effects of text difficulty and readers on predicting reading comprehension from eye movements. In *2015 6th ieee international conference on cognitive infocommunications (coginfocom)* (pp. 407–412).

Davies, M. (2016). *Corpus of Contemporary American English (COCA).* Harvard Dataverse. Retrieved from `https://doi.org/10.7910/DVN/AMUDUW` doi: 10.7910/DVN/AMUDUW

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117–1121.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. doi: 10.1016/j.cognition.2008.07.008

Drieghe, D., Rayner, K., & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 954–969. doi: 10.1037/0096-1523.31.5.954

Duan, Y., & Bicknell, K. (2017). Refixations gather new visual information rationally. In *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 301–306).

Duan, Y., & Bicknell, K. (2020). A rational model of word skipping in reading: ideal integration of visual and linguistic information. *Topics in Cognitive Science*, *12*(1), 387–401.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813. doi: 10.1037/0033-295X.112.4.777

Farid, M., & Grainger, J. (1996). How initial fixation position influences visual word recognition: A comparison of french and arabic. *Brain and Language*, *53*(3), 351–368.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178–210.

González-Garduño, A. V., & Søgaard, A. (2018). Learning to predict readability using eye-movement data from natives and learners. In *Aaai*.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (cmcl 2018)* (pp. 10–18). doi: 10.18653/v1/W18-0102

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, *5*(1), 1–13.

Hollenstein, N., & Zhang, C. (2019). Entity recognition at first sight: Improving ner with eye movement information. *arXiv preprint arXiv:1902.10068*.

Holmes, V., & O'regan, J. (1992). Reading derivationally affixed french words. *Language and Cognitive Processes*, *7*(2), 163–192.

Hyönä, J., Niemi, P., & Underwood, G. (1989). Reading long words embedded in sentences: Informativeness of word halves affects eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(1), 142–152. doi: 10.1037/0096-1523.15.1.142

Inhoff, A., Eiter, B., Radach, R., & Juhasz, B. (2003). Distinct subsystems for the parafoveal processing of spatial and linguistic information during eye fixations in reading. *The Quarterly Journal of Experimental Psychology Section A*, *56*(5), 803–827. doi: 10.1080/02724980244000639

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

Kennedy, A. (2003). The Dundee Corpus [cd-rom]. *Psychology Department, University of Dundee*.

Kennedy, A., Hill, R., & Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th european conference on eye movement.*

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153–168. doi: 10.1016/j.visres.2004.07.037

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Klerke, S., Goldberg, Y., & Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. *arXiv preprint arXiv:1604.03357*.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284. doi: 10.1080/09541440340000213

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: the influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*(1), 12–35.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59.

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Research*, *42*(18), 2219–2234. doi: 10.1016/S0042-6989(02)00131-1

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: an ideal-observer model of reading. *Psychological Review*, *104*(3), 524–553. doi: 10.1037/0033-295X.104.3.524

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 1055–1065).

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22–60. doi: 10.1016/j.cogpsych.2016.06.002

Martínez-Gómez, P., & Aizawa, A. (2014). Recognition of understanding level and language skill using measurements of reading behavior. In *Proceedings of the 19th international conference on intelligent user interfaces* (pp. 95–104).

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. the location of initial eye fixations on words. *Vision research*, *28*(10), 1107–1118.

McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., & Jacobs, A. M. (1989). Eye movement control during reading: II. frequency of refixating a word. *Attention, Perception, & Psychophysics*, *46*(3), 245–253.

Mishra, A., & Bhattacharyya, P. (2018). Predicting readers' sarcasm understandability by modeling gaze behavior. In *Cognitively inspired natural language processing* (pp. 99–115). Springer.

Mishra, A., Dey, K., & Bhattacharyya, P. (2017). Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 377–387).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., …

others (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Morrison, R. E. (1984). Manipulation of stimulus onset delay in reading: evidence for parallel programming of saccades. *Journal of Experimental psychology: Human Perception and performance*, *10*(5), 667.

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal bayesian decision process. *Psychological Review*, *113*(2), 327–357. doi: 10.1037/0033-295X.113.2.327

Okoso, A., Toyama, T., Kunze, K., Folz, J., Liwicki, M., & Kise, K. (2015). Towards extraction of subjective reading incomprehension: Analysis of eye gaze features. In *Proceedings of the 33rd annual acm conference extended abstracts on human factors in computing systems* (pp. 1325–1330).

O'Regan, J. K. (1990). Eye movements and reading. *Reviews of Oculomotor Research*, *4*, 395–453.

O'Regan, J. K. (1992). Optimal viewing position in words and the strategy-tactics theory of eye movements in reading. In *Eye movements and visual cognition* (pp. 333–354). Springer.

O'Regan, J. K., & Lévy-Schoen, A. (1987). Eye movement strategy and tactics in word recognition and reading. In *Attention and performance: Vol. 12. the psychology of reading* (pp. 363–383). Hillsdale, NJ: Erlbaum.

O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugaillère, B. é. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(2), 250.

Perea, M., & Lupker, S. J. (2003). Does jugde activate court? transposed-letter similarity effects in masked associative priming. *Memory & Cognition*, *31*(6), 829–841.

Peters, J., & Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. *Neural networks*, *21*(4), 682–697.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. doi: 10.1037/0033-2909.124.3.372

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific studies of reading*, *10*(3), 241–255.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: a comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(5), 1188–1200. doi: 10.1037/0096-1523.22.5.1188

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. doi: 10.1037/a0020990

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*(4), 504–509. doi: 10.3758/BF03214555

Reichle, E. D., & Laurent, P. A. (2006). Using reinforcement learning to understand the emergence of" intelligent" eye-movement behavior during reading. *Psychological review*, *113*(2), 390.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21. doi: 10.3758/PBR.16.1.1

Sanches, C. L., Augereau, O., & Kise, K. (2017). Using the eye gaze to predict document reading subjective understanding. In *2017 14th iapr international conference on document analysis and recognition (icdar)* (Vol. 8, pp. 28–31).

Sanches, C. L., Augereau, O., & Kise, K. (2018). Estimation of reading subjective understanding based on eye gaze analysis. *PloS one*, *13*(10), e0206213.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in cognitive sciences*, *6*(9), 382–386.

Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, *26*(6), 1270–1281.

Schotter, E. R., Lee, M., Reiderman, M., & Rayner, K. (2015). The effect of contextual constraint on parafoveal processing in reading. *Journal of Memory and Language*, *83*, 118–139. doi: 10.1016/j.jml.2015.04.005

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Slattery, T. J. (2009). Word misperception, the neighbor frequency effect, and the role of sentence context: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1969.

Staub, A., & Goddard, K. (2019). The role of preview validity in predictability and frequency effects on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 110–127. doi: 10.1037/xlm0000561

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Thibadeau, R., Just, M. A., & Carpenter, P. A. (1982). A model of the time course and content of reading. *Cognitive Science*, *6*(2), 157–203.

Vitu, F. (1991). The influence of parafoveal preprocessing and linguistic context on the optimal landing position effect. *Attention, Perception, & Psychophysics*, *50*(1), 58–75.

Von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*(2), 109–127.

Weiss, A. F., Kretzschmar, F., Schlesewsky, M., Bornkessel-Schlesewsky, I., & Staub, A. (2018). Comprehension demands modulate re-reading, but not first-pass reading behavior. *Quarterly Journal of Experimental Psychology*, *71*(1), 198–210.

White, S. J., Rayner, K., & Liversedge, S. P. (2005). The influence of parafoveal word length and contextual constraint on fixation durations and word skipping in reading. *Psychonomic Bulletin & Review*, *12*(3), 466–471. doi: 10.3758/BF03193789

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36. doi: 10.1111/j.1467-9868.2010.00749.x

Yoshimura, K., Kise, K., & Kunze, K. (2015). The eye as the window of the language ability: Estimation of english skills by analyzing eye movement while reading documents. In *2015 13th international conference on document analysis and recognition (icdar)* (pp. 251–255).