UNIVERSITY OF CALIFORNIA, SAN DIEGO

Eye movements in reading as rational behavior

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Linguistics and Cognitive Science

by

Klinton Bicknell

Committee in charge:

Professor Roger Levy, Chair
Professor Jeffrey L. Elman
Professor Andrew Kehler
Professor Keith Rayner
Professor Angela Yu

2011

The dissertation of Klinton Bicknell is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2011

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Roger Levy, to whom I owe a great deal of gratitude. Even beyond his direct contributions to the work reported in this dissertation (on all of which he is a coauthor), Roger substantially improved this work indirectly by teaching me how to be a better academic. As an advisor, Roger excelled at walking the line between always being supportive of all my ideas (even the really crazy ones), while simultaneously giving fantastic feedback, which invariably pushed me into a deeper understanding of whatever I was proposing (and which in turn sometimes helped me to realize that it really was a crazy idea). One of Roger's most salient qualities is that he is never content with a loose understanding of a concept, but constantly strives towards precision and clarity of thought. I am very grateful that through the many meetings and interactions we've had in these past years, Roger has managed to pass this quality on (at least somewhat) to me, and I am sure that my work – and my future career – is substantially enriched because of it.

Special thanks are also due to Jeff Elman, who provided a large amount of advising throughout my graduate career, and who I had the pleasure of collaborating with on research not reported in this dissertation. For the research that is in this dissertation, Jeff contributed a valuable external perspective (and numerous great ideas) on how to make my work compelling for language researchers more broadly, and his contagious excitement always helped me to see new ways in which my work could interest others. In addition, Jeff provided – from the first meetings we had in my first year of graduate school – the amazingly useful service of thinking about my future (even when I wasn't): for example, suggesting I talk to potential post-docs sponsors at conferences early in graduate school, or coming up with suggestions in various meetings for how a certain result would be most effectively presented in a job talk.

I am also very grateful to the rest of my dissertation committee: to Andy Kehler, for advice that was always both pragmatic and inspiring; to Keith Rayner, for never allowing the modeling to become too distant from empirical results; and to Angela Yu, for constantly reminding me of the importance of the details.

During my time in graduate school, I have been fortunate enough to learn the value of collaboration, and to learn it from a number of very knowledgeable and insightful collaborators, who have set high standards for my future collaborations to follow. In addition to those already mentioned, I thank Vera Demberg, Mary Hare, Philip Hofmeister, Marta Kutas, Ken McRae, Emily Morgan, Tim Slattery, and Hannah Rohde. Additionally, I am very grateful to Gordon Legge for sharing the corpus used in the original Mr. Chips experiments, described in Chapter 2.

UC San Diego was a remarkable place to complete a Ph.D. and there are too many people to name who have contributed to the quality of the time I've spent here. Special thanks are due to the technical services of Ezra Van Everbroeck, without whose prompt response to computer crises, a number of conference papers may have included fewer simulation results. Of course, another aspect of this campus that made my time here so valuable is the quality of other students (spanning a range of departments), and I especially thank for their comradery, feedback, and friendship Jamie Alexandre, Rebecca Colavin, Alex Del Giudice, Gabe Doyle, Laura Kertz, Emily Morgan, Bożena Pająk, Albert Park, Hannah Rohde, Lisa Rosenfelt, Liz Schotter, and Nathaniel Smith. UC San Diego also contains a wealth of relevant research groups, and I have been fortunate enough to have contributed to and learned from a number of them: the Computational Psycholinguistics Lab, the Rayner Eyetracking Lab, and the Center for Research in Language, which also provided me with a pre-doctoral fellowship for two years.

Penultimately, I thank my family. I doubt I would have done any of this if it weren't for the continual support of my parents, who have always encouraged and indulged my curiosity and my passions (and also had the foresight to introduce me to computers at a very young age).

Finally, the largest debt is to Bożena, who has been my constant companion on this journey, being there for every deadline, for every success, and for the time in between them. Her unflagging support, her insightful intellect, and her love have touched and enriched every aspect of my work, and of my life.

❖ ❖ ❖

The research was supported by NIH Training Grant T32-DC000041 from the Center for Research in Language at UC San Diego to the dissertation author, and also by a research grant from the UC San Diego Academic Senate, NSF grant 0953870, and NIH grant R01-HD065829 to Roger Levy.

Chapter 2, in full, is an exact copy of the material as it appears in Bicknell and Levy (2010a) [Rational eye movements in reading combining uncertainty about previous words and contextual probability. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1142–1147). Austin, TX: Cognitive Science Society.] 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 84th Annual Meeting of the Linguistic Society of America.

Chapter 3, in full, is an exact copy of the material as it appears in Bicknell and Levy (2010b) [A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL* (pp. 1168–1178). Uppsala, Sweden: Association for Computational Linguistics.] The dissertation author was the primary investigator and author of this paper. In addition to being presented to the Association for Computational Linguistics, this work was also presented at the 23rd Annual CUNY Conference on Human Sentence Processing.

Chapter 4 is the single body chapter not to reproduce a published paper.

Chapter 5, in full, is an exact copy of the material as it appears in Bicknell and Levy (2011) [Why readers regress to previous words: A statistical analysis. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.] 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 24th Annual CUNY Conference on Human Sentence Processing.

VITA

2005	B.A. in Linguistics with Honors University of California, Berkeley
2009	M.A. in Linguistics University of California, San Diego
2011	Ph.D. in Linguistics and Cognitive Science University of California, San Diego

ABSTRACT OF THE DISSERTATION

Eye movements in reading as rational behavior

by

Klinton Bicknell

Doctor of Philosophy in Linguistics and Cognitive Science

University of California, San Diego, 2011

Professor Roger Levy, Chair

Moving one's eyes while reading is one of the most complex everyday tasks humans face. To perform efficiently, readers must make decisions about when and where to move their eyes every 200–300ms. Over the past decades, it has been demonstrated that these fine-grained decisions are influenced by a range of linguistic properties of the text, and measuring eye movements during reading has become one of the primary methods of studying online sentence comprehension. However, it is still largely unclear why linguistic variables affect the eye movement record in the ways they do.

The present work begins to answer this question by presenting a rational framework for understanding eye movement control in reading, in which probabilistic language knowledge plays a crucial role. Specifically, the task of

reading is taken to be one of sentence identification: readers move their eyes to efficiently obtain visual input, which they combine with probabilistic language knowledge through Bayesian inference to yield posterior beliefs about sentence form and structure. Simulations with implemented models within this framework demonstrate that it can provide a principled account of many aspects of reading behavior, including the influence of a number of linguistic variables. In addition, the framework suggests a novel explanation for one of the least understood aspects of eye movements in reading – regressive eye movements – and we present evidence from an eye tracking corpus to support this proposal.

Chapter 1

Introduction

The control of the eyes during reading is one of the most complex learned tasks that humans face everyday. Because human visual acuity is high only in the center of the visual field (the fovea), efficient reading requires moving the eyes (i.e., making a saccade) every 200–300 milliseconds, pushing successive regions of interest into the fovea. To accomplish this, readers must rapidly combine their linguistic knowledge with visual information as well as knowledge of their motor constraints, in order to make decisions about when and where to move the eyes 3–5 times per second. This dissertation seeks to provide new insight into how humans perform this feat – and specifically to elucidate the role played by linguistic factors – through a combination of computational modeling and empirical studies.

Because reading is a prototypical example of a learned skill requiring rapid, complex information processing, gaining a better understanding of how humans control their eyes during reading can also provide insight into the types of solutions that humans find when learning to solve complex learned tasks more generally. Thus, one goal of the present work is to gain insight into eye movements in reading by studying the nature of the solution that humans have found. Specifically – as will be described – we use the tools of *rational analysis* (Anderson, 1990), one framework for studying complex information processing problems, to formalize the demands of the task and to study properties of efficient solutions to it. Comparing these efficient solutions with actual human

behavior can then provide insight into why reading behavior looks the way it does, i.e., how certain aspects of reading behavior may naturally arise as part of efficient reading, given task demands and reader goals.

In addition, while it is well known that eye movements in reading are very responsive to the linguistic properties of the text being read, the precise reasons for the effects of these linguistic variables are still unclear. To the extent that linguistic information plays an important role in the task of reading, our rational analysis will generate proposals for the reasons for these linguistic effects. Further, the existence of linguistic effects has prompted the monitoring of eye movements during reading to become one of the dominant paradigms for researchers studying real-time language processing, despite the unclear reasons that these linguistic effects exist. By precisely articulating the reasons that linguistic effects occur, we also gain a better understanding of the link between language processing and eye movements, which can in turn allow for the eye movement record to be still more informative for language research. These two observations motivate the second goal of the work reported here: to better understand the reasons for effects of linguistic variables on eye movements in reading, which will simultaneously advance our knowledge of why eye movement behavior looks the way it does, and further the state of language processing research studied via eye movements in reading. Prior to introducing our framework, the next two sections elaborate on each of these complementary goals: to better understand aspects of human reading behavior as properties of efficient reading, and to better explain the effects of linguistic variables on eye movements.

1.1 A rational account of a complex behavior

To better understand aspects of human reading behavior as properties of efficient reading, we follow a research paradigm for analyzing complex information processing tasks called rational analysis introduced by Anderson (1990). In rational analysis, the researcher first formalizes the goals and the cognitive and physical constraints relevant to the task, as well as the information avail-

able to the agent, and then develops a model of optimal behavior under those conditions. To the extent that the behavior of the model is similar to that of humans, this provides a new way of understanding the reason why human behavior looks the way it does – it is an efficient way to solve the problem. In this case, we can also learn a lot about various aspects of human behavior by understanding the formal structure of the optimal solution. This paradigm is thus ideally suited to generate proposals for why human reading behavior looks the way it does as well as to evaluate the efficiency of the solutions to the problem that human readers have found.

Rational analysis has been perhaps most successfully applied to relatively low-level sensory and motor tasks (e.g., Weiss, Simoncelli, & Adelson, 2002; Körding & Wolpert, 2004; Trommershäuser, Maloney, & Landy, 2008), in which the fits obtained between the optimal model and human data can be very close. These results may be taken as unsurprising, as an argument can easily be made that evolution has had a very long time to optimize these low-level sensorimotor systems, which have existed for a large part of our evolutionary history. Additionally, the paradigm has been successfully applied to problems in higher level cognition, such as finding one's way through a complexly structured environment (Stankiewicz, Legge, Mansfield, & Schlicht, 2006), object perception (Kersten, Mamassian, & Yuille, 2004), and visual search (Najemnik & Geisler, 2005, 2008), results which might be taken to suggest that the human brain can optimize its performance at a relatively broad range of tasks. However, while these abilities are not quite as old as sensorimotor systems, they still represent skills that evolution has very plausibly had ample time to optimize, and thus, these results may also be the result of evolutionary optimization. From this perspective, the case of reading is particularly interesting: since written language has only existed for thousands of years, evolution would not have plausibly had time to optimize performance at it. Thus, to the extent that humans approximate optimal performance, it would provide a new source of evidence that the human brain can implicitly find optimal solutions to challenging, new problems. Because reading is such a highly practiced skill (in industrialized societies), it represents one of the best chances (among other complex learned skills) of hu-

mans having sufficient time and motivation to find the most efficient solution.

1.2 Reasons for effects of linguistic variables

The second goal of this work is to explain effects of linguistic variables on eye movements in reading. In the past decades, the eye movement control literature has richly documented effects of many linguistic properties of the text (see Rayner, 1998, 2009 for reviews). More recently, sophisticated computational models of eye movements in reading have been developed, which give formal accounts for many of these effects. Among the most well known of these models are *E-Z* Reader (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Warren, & McConnell, 2009) and SWIFT (Engbert, Longtin, & Kliegl, 2002; Engbert, Nuthmann, Richter, & Kliegl, 2005). However, in these models, linguistic effects have typically been modeled as variables exogenous to the system. As one example, in *E-Z* Reader, the time taken to complete the first stage of the processing of a word with frequency f and in-context predictability p is said to be distributed according a Gamma distribution with a mean given by

$$t(L_1) = \alpha_1 - \alpha_2 \log f - \alpha_3 p \tag{1.1}$$

where α_1, α_2, and α_3 are free parameters fit to the data.[1] The model thus clearly predicts that frequency and predictability should have effects on word processing times, and that these effects should be linear, yet it does not give any insight into why these relationships hold. Here, by providing a rational analysis of eye movements in reading, we investigate to what extent we can derive the form of relationships such as that in Eq. 1.1 from the structure of the problem and principles of rational inference.

[1] This oversimplifies slightly. In fact, the value resulting from Eq. 1.1 is then further adjusted by the model to incorporate effects of the word's length and distance from the fovea.

1.3 A rational framework for reading

In order to achieve these two goals, we propose a rational framework for understanding eye movements in reading, which this section introduces. Specifically, we formalize the problem of reading as one of identifying the text being read in an efficient manner.[2] There are exactly two sources of information relevant to the identification of the text. Perhaps the most obvious is visual input, at least some of which is generally required to be confident in the identity of the text. The other is knowledge of the language in which the text is written. In addition to providing the relatively trivial information about what words exist in the language, language knowledge can provide contextual information about what words are more or less likely in context. As a rather extreme example, when reading the sentence 'The children went outside to . . . ', not much visual input is necessary to be confident that the next word is *play*.

Thus, both language knowledge and visual input provide information about the identity of the text. The normative way to combine these two sources of information is through Bayesian inference, where language knowledge provides the prior beliefs about what sentences are more or less likely, and then those beliefs are updated with the visual information, which provides the likelihood, to yield posterior beliefs about the identity of the text given both sources of information. Since language knowledge is possessed prior to reading, the task of reading becomes one of deciding when and where to move the eyes to efficiently gather visual information, given the language knowledge. Thus, the ultimate reason for effects of linguistic properties of the text on eye movement behavior given by this framework is that those linguistic properties change the amount of visual input that is required to become confident about the identity of a particular part of the text. We next highlight two important properties of this framework before introducing the contents of the body chapters of this dissertation.

[2] Of course, identifying the text is clearly not all that a reader is doing, but all reader goals are achieved via the identification of some parts of the text, so we consider it a reasonable approximation.

1.3.1 Uncertainty

Because the goal of reading is taken to be one of efficient identification, that entails that readers will have some trade-off between speed and accuracy. For all but the lowest values placed on speed relative to accuracy, efficient reading will necessarily mean that a reader will sometimes be relatively less confident about the identity of some parts of a text; e.g., readers will have non-negligible amounts of uncertainty about the identity of some words. It is apparent that the amount of uncertainty that a reader in this framework has about a particular word depends on both how much visual input is obtained about that word as well as the preceding context. Crucially, in addition, the level of uncertainty about that word also depends on its following context. Thus, the uncertainty about a word will necessarily change over time. This fluctuating level of certainty about the identities of previous words is a natural consequence of reading in this framework, and factors heavily into the framework's explanation for regressive eye movements discussed below.

1.3.2 Integrating word identification and reading

One criticism commonly leveled against models of eye movements in reading such as *E-Z* Reader and SWIFT is that they fail to make use of many insights that have been gained from the study of single word identification (see, e.g., the discussion in Reichle, Rayner, & Pollatsek, 2003). Specifically, in order to make use of all the results obtained in the single word identification literature, the argument is that it would be beneficial to have a single model of eye movements in reading which incorporates within it a detailed model of single word identification. From one perspective, the framework and models described in this dissertation satisfy this goal: we are explicitly modeling the identification of the words in a text from visual input, and thus should obtain standard effects produced by models of single word identification (which are not necessarily encoded in current models of eye movements in reading) such as effects of visual neighborhood. From this perspective, the work presented here can be seen as an extension of Norris's (2006, 2009) Bayesian Reader model of single word iden-

tification. From another perspective, however, the framework presented here is not simply a model that incorporates within it a model of single word identification. As will be demonstrated, because the task is taken not to be one of serial word identification, but rather one of efficient text identification, the task differs in a number of important respects from that of single word identification.

1.4 Chapter introductions

The body chapters of this dissertation further develop this rational analysis by discussing two models within the rational framework and showing that they can both give insight into known effects and make new predictions for a less studied effect, between-word regressions. At a broad level, the first two body chapters describe models and the other two evaluate model predictions. Specifically, Chapter 2 describes an extension to a previous rational model of eye movements in reading and Chapter 3 presents a new model in the framework. Chapter 4 tests predictions of the new model for the well-known linguistic effects of a word's in-context predictability, frequency, and length, and Chapter 5 tests the predictions of the new model for regressive saccades.

With the exception of Chapter 4, the body chapters of this dissertation are exact copies of published papers, with their own introductions and conclusions, and thus each can stand on its own and can be read independently of the others. In order to highlight the close relationships between them and to join them together into a coherent argument, the remainder of this section briefly introduces and situates the contents of each.

1.4.1 Chapter 2

Chapter 2, in full, is an exact copy of the material as it appears in Bicknell and Levy (2010a) [Rational eye movements in reading combining uncertainty about previous words and contextual probability. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1142–1147). Austin, TX: Cognitive Science Society.] It describes an

extension to a previous rational model of eye movements in reading, called Mr. Chips (Legge, Klitz, & Tjan, 1997; Legge, Hooven, Klitz, Mansfield, & Tjan, 2002). Mr. Chips is a model that fits within our rational framework, but using very simplistic models of visual input and language knowledge. Visual input in Mr. Chips is simply the veridical identities of nine characters around the point of fixation (as well as word boundary information extending further), and the model of language knowledge is only information about the frequency with which each word in the language occurs (i.e., a unigram model). Thus, Mr. Chips cannot make use of linguistic context. Given these two information sources, Mr. Chips then combines them together through normative Bayesian inference to form a posterior distribution over the text being read, and uses that posterior distribution to decide on its next action.

Mr. Chips is a model only of where to move the eyes, and makes no predictions for fixation durations. The model identifies words serially, from left to right, and uses a very simple algorithm to select the target of the next saccade. Specifically, it aims the next saccade towards the position that will, on average, provide the most information about the word it is currently trying to identify. It continues to select saccades in this manner until the current word is identified with 100% certainty, and then begins to do the same for the next word. Despite using this very simple heuristic for targeting saccades, the model produces behavior that replicates a number of human findings in word skipping rates, initial fixation locations on words, and refixation rates. The fact that such an overly simplistic model can reproduce a range of human reading phenomena suggests that many properties of human eye movements in reading may indeed arise from efficient solutions to the problem.

In the work presented Chapter 2, we extend the model in two ways, and show that the resulting model more closely matches human performance on two measures. Specifically, we allow the model to maintain some uncertainty about the identities of previous words and allow it to make use of linguistic context in identifying words, arguing that there is evidence that humans do each of these. We present simulations run with the resulting model that show that its average saccade size and skipping rates are closer to those of human readers. This is thus

a case in which adding knowledge that humans are likely to have to a rational model makes its predictions closer to human data, further supporting the notion that humans may have found a very efficient solution to the problem of reading.

This approach does have a number of limitations, however. Perhaps most importantly, neither the Mr. Chips model nor our extension of it can model fixation durations, which are one of the most commonly studied facets of eye movements in reading – and the primary locus of many linguistic effects. The main reason for this limitation is the way that visual input is obtained. Recall that visual input in Mr. Chips consists of the veridical identities of nine characters centered around the point of fixation. Thus, after spending a single model timestep fixating a particular position, the model already has veridical knowledge of the nine characters around it – and will not gain any more knowledge by fixating that position longer. One strategy to overcome this limitation would be to replace the veridical visual input with 'noisy' (or stochastic) visual input, which gives only partial information about the identities of the characters around the point of fixation. In this way, the model may spend multiple timesteps fixating the same position in order to get more visual input about that region. The work presented in the following chapter describes a new rational model of eye movements in reading which makes use of exactly this strategy in order to make predictions for fixation durations.

1.4.2 Chapter 3

Chapter 3, in full, is an exact copy of the material as it appears in Bicknell and Levy (2010b) [A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL* (pp. 1168–1178). Uppsala, Sweden: Association for Computational Linguistics.] It presents our rational framework for analyzing eye movements in reading and, in addition, describes a new model in this framework that improves upon the Mr. Chips model in a number of ways. Like the extension of Mr. Chips described in the previous chapter, this model does make use of linguistic context to help identify words. Unlike the extension of Mr. Chips, however, visual

input in the model is not veridical letter identities, but rather noisy probabilistic information. Visual acuity is explicitly encoded in the model as the level of noise for the information about a particular character depending on its position in the visual field. The use of noisy visual input means that a reader must decide how many timesteps to spend fixating a particular location, and thus, unlike the Mr. Chips extension, this model does make predictions for fixation durations.

Abstracting over the computational details, the inference method is the same as in Mr. Chips: language knowledge is combined with visual information using Bayesian inference to yield a posterior distribution over possible identities of the sentence. The way in which this model selects its actions, however, is different. Recall that the Mr. Chips model focused on identifying words serially, and selected its next action to be the one that would be expected to give the most information about the identity of the current word of interest. Note that if the ultimate goal is to identify an entire sentence efficiently, then such a strategy can be shortsighted and suboptimal. For example, imagine a situation in which some action A will provide the most information about the current word of interest, but some action B will provide almost as much information as A and also substantially more information about the following words as well. The Mr. Chips model will select action A, which may result in taking a longer time to read the full sentence. To avoid this problem, for the new model presented in this chapter, we use machine learning techniques to approximate globally optimal behavior, which will result in selecting actions to read full sentences as efficiently as possible.

In addition to presenting the rational framework and model, this chapter also uses the framework to explore the relationship of regressive saccades to efficient reading behavior. As described above in Section 1.3.1, one problem that a reader in the rational framework must contend with is that their confidence in the identity of a previous region of a sentence will sometimes fall as a result of information further downstream. Imagine that a reader has a strategy of moving the eyes generally left-to-right[3] and bringing confidence in each word up to some desired level. If the reader simply ignores this problem, then

[3]Without loss of generality, we assume a left-to-right language.

their representation of the content of a word for which confidence falls due to subsequent material will be lower quality than originally intended. In this chapter, we propose two possible ways in which a reader can alleviate this problem. The first is to simply slow down, and gather more visual input about each part of the sentence, which will result in levels of confidence higher than the actual desired level. If these levels of confidence are high enough, then there will be little chance of confidence falling below the desired level, even if it does fall somewhat. The second strategy is to make a regressive eye movement when the confidence about a previous region falls below a certain value. We use the implemented model to explicitly compare these two strategies and show that for a range of possible reader goals (i.e., relative values of speed vs. accuracy) the strategies that use regressive eye movements result in more efficient reading than simply slowing down.

This finding is important for at least two reasons. First, it changes the way we look at regressive eye movements. Since at least Buswell (1920), regressive eye movements in reading have been viewed solely as a sign of bad reading behavior, but this result suggests that they are an important device which allows skilled readers to read more efficiently. Second, the occurrence of regressions to previous regions is a phenomenon that models of reading such as *E-Z* Reader cannot easily handle. Since they assume that readers serially identify each word in the sentence, there is no reason within the model that a reader should return to a previous word. In fact, the latest version of *E-Z* Reader is an attempt to make predictions for regressive saccades, but it must do so by adding another exogenous function ('integration failure') to the model. Conversely, in our rational framework regressions are a natural response to confidence falling about previous regions, a problem posed to readers that is integral to our model. Note, however, that while this chapter proposes a new reason why readers should produce regressive eye movements, it does not present any evidence that readers actually do make them in this situation. Chapter 5 presents human data that support of this claim.

1.4.3 Chapter 4

Chapter 4 is the single body chapter not to reproduce a published paper. It is the first of two chapters which compare predictions of the new model described in Chapter 3 and is the only chapter in this dissertation that presents unpublished work. Specifically, it explores the predictions of the new model for effects on eye movements of three linguistic variables: word predictability, frequency, and length. The first part of the chapter presents intuitions for why the model's behavior might be expected to display effects of these three variables, and what these effects should be expected to look like. We give clear intuitions from the model for why words of high predictability, high frequency, and short length should be fixated fewer times and for less time, as is the case for human eye movements. In addition, however, we note that some technical limitations of the current implementation of the model may cause an inverse length effect: longer words having fewer, shorter fixations.

The chapter also reports simulations run with the model to analyze the effects of these three variables that the model actually produces. As predicted given the intuitions described above, the model does produce shorter, fewer fixations on words that are highly predictable and highly frequent. As may be anticipated, however, the model's predictions for effects of word length are less clear: the effects are in the predicted direction for words of length up to four characters, but past that, effects of length are either in the wrong direction or absent completely, suggesting that the aforementioned limitations of the model do indeed impede its ability to produce human-like length effects. A second set of simulations is also presented, which shows that a modified version of the model for which one of the limitations is removed produces length effects that are more like those of human readers.

Thus, this chapter demonstrates that the new model can successfully account for the most well-known linguistic effects on eye movements in reading: word predictability, frequency, and – to a limited extent – word length. The less robust ability of the model to account for effects of length highlight some of the current limitations of the framework, as well as obvious directions forward.

1.4.4 Chapter 5

Chapter 5, in full, is an exact copy of the material as it appears in Bicknell and Levy (2011) [Why readers regress to previous words: A statistical analysis. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.] It reports empirical analyses of regressive saccades made by human readers in natural reading, aimed at testing the new model's predictions for regressive saccades. We formalize the predictions made by the new model's account of between-word regressions described in Chapter 3 – namely, that readers make regressions when their confidence about the identity of a previous region falls – and compare it with predictions made by a range of other accounts of between-word regressions. Specifically, we test the ability of these accounts to predict when people will make a regression back to the previous word in a large corpus of eye movements produced by readers of newspaper editorials (Kennedy & Pynte, 2005).

The results of the analysis reveal that the confidence falling account correctly predicts the full range of linguistic effects observed in the dataset. Some other accounts can predict some of these effects, but the confidence falling account is the only single model to predict them all. We note, however, that an account in which regressions are made for multiple reasons, effectively combining other accounts, could also explain the full pattern of results. These results thus provide further support to the explanation of regressions made by our framework, and in so doing also provide further evidence that many aspects of human reading behavior are well modeled as resulting from efficient solutions to the problem of reading.

1.4.5 Chapter 6

Chapter 6 concludes the dissertation. It reviews the arguments made and evidence presented in this dissertation in favor of understanding many aspects of human eye movements in reading as naturally resulting from efficient reading behavior. In addition, it points to current and future directions for this line

of inquiry.

Chapter 2

Rational eye movements in reading combining uncertainty about previous words with contextual probability

Klinton Bicknell and Roger Levy

Abstract. While there exist a range of sophisticated models of eye movements in reading, it remains an open question to what extent human eye movement behavior during reading is adaptive given the demands of the task. In this paper, we help to answer this question by presenting a model of reading that corrects two problems with a rational model of the task, Mr. Chips (Legge et al., 1997). We show that the resulting model is closer to human performance across two measures, supporting the idea that many components of eye movement behavior in reading can be well understood as a rational response to the demands of the task.

2.1 Introduction

Choosing when and where to move one's eyes during reading is one of the most complicated skilled tasks humans perform. While there are a number of computational models achieving good numerical fits on eye movement data from reading (e.g., Reichle, Pollatsek, & Rayner, 2006; Engbert et al., 2005), it is still unclear to what extent the complex behaviors observed are rational responses to the demands of the problem itself and to what extent they arise from the idiosyncrasies and restrictions of human cognition. Legge, Klitz, and Tjan (1997) started to answer this question with Mr. Chips, a model which predicts eye movements that approximate an optimal solution to one formalization of the task of reading. Legge et al. pointed out that their model's behavior exhibits a number of patterns also found in human reading, providing evidence for understanding those behaviors as rational responses to the task. Despite its success, however, the Mr. Chips model oversimplifies two important aspects of the problem of reading, and also has empirical problems accounting for human reading behavior in two domains. In this paper, we propose a model extending Mr. Chips that removes these two oversimplifications to make the model's task more similar to that faced by humans. We show that the resulting model also remedies the two empirical deficiencies in Mr. Chips, further supporting the notion that many aspects of human reading behavior can be explained as rational responses to the demands of reading.

The essentials of the problem of making eye movements in reading are determining how long to leave the eyes in a given spot and – when a reader decides to move them – where to go. These decisions are made sequentially to produce the alternating sequence of fixations (relatively stable periods) and saccades (movements) that characterizes the eye movement record. The past 30 years have seen a proliferation of experimental studies investigating this topic, which have answered a number of low-level questions such as the nature of the perceptual span and constraints on saccade latency as well as questions concerning the relationship between eye movements and higher-level cognitive processes such as the effect of word frequency and predictability (see Rayner,

1998 for an overview). Sophisticated computational models have been developed based on these findings, the most well-known of which are E-Z Reader (Reichle et al., 1998, 2006) and SWIFT (Engbert et al., 2005). Both E-Z Reader and SWIFT assume that lexical processing (or word recognition) is the primary driver for eye-movements in reading, and both have enjoyed considerable success, in large part because they achieve very good fits to eye movement data from reading in a number of contexts, using a relatively small number of parameters. Despite their empirical strength, they fail to illuminate the reason why human reading behavior looks the way it does in one crucial respect – the extent to which it resembles a rational response to the problem posed by reading.

One leading approach for answering such questions is that of rational analysis (Anderson, 1990), a paradigm in which one formalizes the goals and cognitive and physical constraints relevant to a problem and develops a model of optimal behavior under those condition. To the extent that the behavior of the model is similar to that of humans, this provides a new way of understanding the reason why human behavior looks the way it does – it is the best way to solve the problem. The relationship between rational models and models such as E-Z Reader and SWIFT is well understood in terms of Marr's (1982) levels of analysis. Marr distinguishes three levels of mutually-constraining analyses that can be performed on cognitive processes: the *computational* level, which specifies the nature of the computation being performed, the information relevant to solving it, and the way to combine that information to solve it; the *algorithmic* level, which specifies the representation for the input and output and the algorithm by which the agent goes about solving it; and the *implementational* level, which specifies how the representations and algorithm are realized neurally. In these terms, rational models generally provide answers at the computational level of analysis. Models such as E-Z Reader and SWIFT help us to understand the algorithmic level, but cannot answer questions about the extent to which human reading is rational.

Legge et al. (1997) presented a computational level analysis of reading, formalizing the central task – as in E-Z Reader and SWIFT – as one of serial word identification. They presented the Mr. Chips model, which approximates

optimal behavior under their formalization, and shows a number of similarities with human reading behavior. Here, we point out two problems with their model of reading. First, their model takes the task to be to identify a string of independent words rather than a coherent sequence, i.e., their model does not make use of linguistic context, which experimental work suggests that humans use (McDonald & Shillcock, 2003). Second, it assumes that the task of the reader is to identify each word with complete certainty, yet recent evidence suggests that readers maintain uncertainty as to the identities of previous words (Levy, Bicknell, Slattery, & Rayner, 2009). In addition to these problems in their model's design, the model also makes incorrect predictions for two relatively basic measures of eye movements in reading: saccade sizes and word skipping rates. The model we present fixes these two design problems by including linguistic context and using a flexible word identification criterion, and results in improved performance in accounting for human saccade sizes and word skipping rates.

The plan of the remainder of the paper is as follows. First, we describe the details of the Mr. Chips model, along with its empirical successes and failures. Next, we describe our extension of the Mr. Chips model, and finally present two experiments showing that fixing each of the two design problems results in performance more like humans.

2.2 Mr. Chips

The task of reading in the Mr. Chips model (Legge et al., 1997) is one of planning saccades for serial word identification. That is, the model works by gathering visual input from the current fixation location and using that visual input to plan a saccade. That saccade is then executed (with some motor error), visual input is obtained from the new location, and the cycle repeats. When one word is identified with 100% confidence, identification of the next word begins. Thus, the only decision the model makes is where to move the eyes next. There are just three sources of information relevant to making that decision. Visual input and knowledge of the language are combined to identify words, and knowledge of the motor error in the system assists in the planning problem.

Since it forms the basis for our model, we describe the Mr. Chips model here in detail, discussing in turn each of the sources of information and then the algorithm by which the model combines them to choose saccades. To match the description of our model later, we use a notation a bit different than Legge et al. to describe Mr. Chips.

2.2.1 Information sources

Visual input

The visual input in Mr. Chips consists of the veridical identities of the nine characters centered on the fixated character (representing the visual fovea), as well as partial information about the four characters on either side of this range (representing the visual periphery). This partial information is simply word boundary information, indicating whether each character is part of a word or not (e.g., a space). The number of characters in each of these ranges was chosen to be representative of the perceptual span for readers of English, known to be around 17–19 characters (Rayner, 1998).

Language knowledge

The model's knowledge of language consists of simply word frequency information, i.e., a unigram model. Note that this means the model cannot make use of the linguistic context to aid in word identification.

Motor error

The final component of the model's knowledge of the task is that of motor error, the distribution of a saccade's landing position given the intended target position the model chooses. In Mr. Chips, the ith landing position ℓ_i is normally distributed around the ith intended target position t_i with a standard deviation

of 30% of the intended distance[1]

$$\ell_i \sim \mathcal{N}\left(t_i, (0.3 \cdot |t_i - \ell_{i-1}|)^2\right). \tag{2.1}$$

2.2.2 Model

We now give the algorithm that the Mr. Chips model uses to select the intended target for the next saccade. First, note that given the visual input obtained by the model from the first to the ith fixation $\mathcal{I}_1^i$ and the word frequency information, the model can calculate the posterior probability of any possible identity of a word w that is consistent with the visual input by normalizing its probability from the language model by the total probability of all visually consistent identities,

$$p(w|\mathcal{I}_1^i) = \frac{\chi(\mathcal{I}_1^i, w)p(w)}{\sum_{w'} \chi(\mathcal{I}_1^i, w')p(w')} \tag{2.2}$$

where $\chi(\mathcal{I}, w)$ is an indicator function with a value of 1 if w is consistent with the visual input $\mathcal{I}$ and 0 otherwise, and $p(w)$ is the probability of w under the language model.

To identify a given word, the model selects the saccade target $\hat{t}_i$ that, on average, will minimize the entropy in this distribution, i.e., that is expected to give the most information about the word's identity

$$\hat{t}_i = \operatorname*{argmin}_{t_i} E\left[H(w|\mathcal{I}_1^i)|t_i, \mathcal{I}_1^{i-1}\right] \tag{2.3}$$

$$= \operatorname*{argmin}_{t_i} \sum_{\mathcal{I}_i} H(w|\mathcal{I}_1^i)p(\mathcal{I}_i|t_i, \mathcal{I}_1^{i-1}). \tag{2.4}$$

That is, the minimum can be found by calculating the conditional entropy produced by each possible new input sequence and weighting those entropies by the probability of getting that input sequence given a choice of target location. In information theory (Cover & Thomas, 2006), conditional entropy is standardly

[1] In the terminology of the literature, this model has only 'random' motor error (variance) and not 'systematic' motor error (bias), under the assumption that an optimal model would just compensate for any systematic problems with its motor control system.

defined as

$$H(w|\mathcal{I}_1^i) = -\sum_w p(w|\mathcal{I}_1^i) \log p(w|\mathcal{I}_1^i). \tag{2.5}$$

The second term in the formula for $\hat{t}_i$, the probability of a particular visual input given a target location and previous input, is given by marginalizing over possible landing positions

$$p(\mathcal{I}_i|t_i, \mathcal{I}_1^{i-1}) = \sum_{\ell_i} p(\ell_i|t_i) p(\mathcal{I}_i|\ell_i, \mathcal{I}_1^{i-1}) \tag{2.6}$$

and then possible words

$$p(\mathcal{I}_i|\ell_i, \mathcal{I}_1^{i-1}) = \sum_w p(\mathcal{I}_i|\ell_i, w) p(w|\mathcal{I}_1^{i-1}). \tag{2.7}$$

Putting these together, we have that $\hat{t}_i$ is selected as

$$\operatorname*{argmin}_{t_i} \sum_{\mathcal{I}_i} H(w|\mathcal{I}_1^i) \sum_{\ell_i} p(\ell_i|t_i) \sum_w p(\mathcal{I}_i|\ell_i, w) p(w|\mathcal{I}_1^{i-1}). \tag{2.8}$$

That is, we can calculate the expected conditional entropy for each possible value of t_i by summing over all possible inputs, whose probabilities are given by summing over all possible identities of the word and landing positions. To see that this sum ranges over a finite number of values, note first that there are only a finite number of possible word identities w to sum over. Given the possible word identities, there are only a finite number of landing positions ℓ_i for which the visual information could possibly help in identifying the word – any landing positions outside this range will not produce any reduction in entropy. Since there is a single visual input $\mathcal{I}_i$ for each combination of landing position and word identity, this summation is over a finite range. To ensure finiteness of the search to find the value of t_i that produces the minimum entropy, Mr. Chips only searches those within the range of the ℓ_i that could give some information about the current word. In case of ties, the model selects the furthest position to the right.

2.2.3 Comparing Mr. Chips to humans

Legge et al. (2002) present a number of ways in which the behavior of the Mr. Chips model is similar to human reading behavior. The model produces be-

havior that replicates a number of human findings in word skipping rates, initial fixation locations on words, and refixation rates. The result for word skipping rates – where word skipping for the model is defined as never having any of the word's characters as the centrally fixated character – is that longer words are skipped less often, and the slope of the relationship between word length and skipping rate has a very similar slope for the model as for humans. For initial word fixation locations, or landing positions, the model replicates the human behavior of most commonly landing at or just to the left of the word's center, and also the fact that the landing position shifts toward the left as the launch site of the saccade shifts further to the left. For refixations, the model mimics human behavior in showing the proportion of refixations to increase with word length, and in addition, within a given word length class, the model refixates low frequency words a higher proportion of the time than high frequency ones. Finally, as a function of landing position, refixations are the least likely for the model, as for humans, when the initial landing position is near the center of the word.

As noted above, however, it is also the case that the model exhibits some behavior very different from that of human readers. For example, the model's average saccade length is just 6.3 characters, noticeably lower than that for humans, who are around 8 (Rayner, 1998). Second, although, as mentioned, the slope of the relationship between word skipping rates and word length has a similar slope for the model as for humans, the model skips far fewer words than humans do.[2] In short, judging by these two measures, a rational model that is using all the information available and expensively calculating the best saccades to reduce entropy in word identification appears to be reading slower than humans do.

In rational analysis, the fact that an 'optimal' model is performing worse than humans (here in terms of speed) suggests two likely problems: (a) the model is not making use of all the information that humans use or (b) the

[2]The graph given in Legge et al. (2002) appears to show remarkably similar word skipping rates between the model and humans, but that graph is from the sole simulation in the paper for which Legge et al. assumed no motor error. When motor error is included, the skipping rates are significantly lower for the model than for humans, as shown in Figure 2.1.

model's computational goal is not the same as the one that humans are solving. As suggested above, we argue that in this case both reasons are partially to blame. Since it has only word frequency information as its model of language, the Mr. Chips model cannot make use of linguistic context to aid in word identification, while there is evidence that humans make heavy use of it. The model also assumes that the goal is to identify each word with 100% confidence, but experiments suggest that humans do not. In the next section, we modify the Mr. Chips model to include some information about linguistic context and a flexible identification confidence criterion.

2.3 Extending Mr. Chips

The model described here generalizes the Mr. Chips model in three ways. First, it can use an arbitrary language model as its source of language knowledge, and thus make use of information about the linguistic context in word identification, solving the first problem with Mr. Chips we pointed out above. Second, it can move on to the next word after it achieves a flexible level of certainty about the current word's identity, solving the second problem. Finally, our model also allows for the standard deviation of the motor error to be an arbitrary linear function of the intended distance, allowing us to incorporate a more realistic motor error function. We describe the model in the same format as we described the Mr. Chips model, first describing its sources of information, and then its algorithm for selecting saccade targets.

2.3.1 Information sources

Visual input

The visual input component is unchanged from the original Mr. Chips model.

Language knowledge

The model's knowledge of language is represented by an arbitrary language model that can generate string prefix probabilities, e.g., an n-gram model or a probabilistic context-free grammar (*PCFG*). Such models can capture the between-word dependencies needed for the model to make use of linguistic context in word identification.

Motor error

In our model as in Mr. Chips, the ith landing position is normally distributed around the ith target location, except that the standard deviation is an arbitrary linear function of the intended distance

$$\ell_i \sim \mathcal{N}\left(t_i, (\beta_0 + \beta_1 |t_i - \ell_{i-1}|)^2\right) \tag{2.9}$$

allowing for the use of a more realistic motor error function. Experiments in this paper use the one from SWIFT (Engbert et al., 2005; $\beta_0 = 0.87$, $\beta_1 = 0.084$).

2.3.2 Algorithm

As in the original Mr. Chips model, at any given point in time, the model is working to identify one word. However, this revised model considers the goal of identifying this word achieved when the marginal probability of some identity for the word given the visual input exceeds a predefined threshold probability α. This flexibility requires the algorithm to be substantially modified to allow for uncertainty about previous word identities and the use of linguistic context. We denote the sequence of words as W, where the first word is W_1.

Because every word in Mr. Chips was identified with complete certainty, the model always knew precisely at which position the next word to be identified began, and its goal was always to identify this next word. Now that the model has uncertainty about the identities of previous words, however, the goal must be changed. In the revised model, the reader is always focused on some character position n, and its goal is to identify whether some word $W_{(n)}$ begins at that position, and if so, which one, with confidence exceeding α. Once the

model has achieved this goal, it then chooses a new character position n via a procedure whose description we leave for later. To be explicit about this goal, we slightly update our original equation for choosing $\hat{t}_i$, swapping w out for $W_{(n)}$,

$$\hat{t}_i = \underset{t_i}{\operatorname{argmin}} \sum_{\mathcal{I}_i} H(W_{(n)}|\mathcal{I}_1^i) p(\mathcal{I}_i|t_i, \mathcal{I}_1^{i-1}) \tag{2.10}$$

where the conditional entropy is calculated assuming that some word does in fact begin at position n. The fact that our language model can now make use of linguistic context means that the equation for finding the probability of the current word given some visual input (Equation 2.2) must also be changed to marginalize over identities of the preceding words

$$p(W_{(n)}|\mathcal{I}_1^i) = \sum_{W_1^{(n)-1}} p(W_{(n)}|\mathcal{I}_1^i, W_1^{(n)-1}) p(W_1^{(n)-1}|\mathcal{I}_1^i). \tag{2.11}$$

These probabilities of strings consistent with the visual input are again given probabilities according to their probability in the language model normalized by the probability of all other consistent strings (cf. Equation 2.2)

$$p(W|\mathcal{I}_1^i) = \frac{\chi(\mathcal{I}_1^i, W) p(W)}{\sum_W \chi(\mathcal{I}_1^i, W) p(W)}. \tag{2.12}$$

The second term in Equation 2.10 is expanded as in Mr. Chips by marginalizing over possible landing positions

$$p(\mathcal{I}_i|t_i, \mathcal{I}_1^{i-1}) = \sum_{\ell_i} p(\ell_i|t_i) p(\mathcal{I}_i|\ell_i, \mathcal{I}_1^{i-1}), \tag{2.13}$$

but now to incorporate information about the linguistic context, we must next marginalize over possible full sentence strings instead of possible words

$$p(\mathcal{I}_i|\ell_i, \mathcal{I}_1^{i-1}) = \sum_W p(\mathcal{I}_i|\ell_i, W) p(W|\mathcal{I}_1^{i-1}). \tag{2.14}$$

If we make the simplifying assumption that the model does not consider possible future input about words that are after $W_{(n)}$, this sum can again be finitely computed for a given t_i by a relatively straightforward dynamic programming

scheme. The range of possible values of t_i to search through also grows relative to Mr. Chips, because the model must consider not only any position that can give visual input about $W_{(n)}$ itself, but also positions that can give information about any position of uncertainty, since that may indirectly help to identify $W_{(n)}$ through linguistic context. In the case where the language model is an n-gram model, it can be shown that the minimum value of t_i that can contribute toward helping to identify $W_{(n)}$ only extends back to the first uncertain character after the most recent string of $n-1$ words for which the model has no residual uncertainty. Having established the method of selecting a saccade to identify $W_{(n)}$, we next give a description of the full algorithm of the model, including how to select n.

The model always begins reading by focusing on identifying $W_{(0)}$. Once the probability of some identity for $W_{(0)}$ is greater than α, all the possible identities of $W_{(0)}$ that have not been ruled out by visual input are combined into a set of possible 'prefixes'. Each of these prefixes has a conditional probability given the visual input, and each one predicts that the next word in the sentence begins at a particular position (i.e., two characters past the end of that string). Thus, the set of prefixes specify a probability distribution over the possible positions at which the next word begins. The model simply selects the most likely such position as the next character position n to focus on identifying $W_{(n)}$.

Now in the general case, the system has a set of prefixes together with their conditional probabilities given the visual input, and a position n, which it is trying to identity the word beginning at. It plans and executes saccades according to the formula for $\hat{t}_i$, and after getting each new piece of visual information, the model rules out not only possible candidates for the current word, but also possible prefix strings, and renormalizes both distributions. The model's attempt to identify $W_{(n)}$ can now end in one of two ways: (a) the model's confidence in some identity of $W_{(n)}$ exceeds the confidence threshold α or (b) the model eliminates all possible candidates for $W_{(n)}$ and thus knows that no word begins at that position. In the former case, the model creates all possible concatenations of prefixes ending 2 characters prior to $W_{(n)}$ (i.e., prefixes whose next word begins at n) with all the possible identities of $W_{(n)}$, and adds these

new strings to the set of prefixes. Then, in both cases, it removes those original prefixes whose next word begins at n from the set. Note that this update of the list of prefixes leaves unaffected prefixes that are incompatible with a word beginning at position n, but still compatible with visual input. Finally, since the set of prefixes again gives a distribution over the starting position of the next word, the model selects the most likely new n and the cycle continues.

2.4 Experiment 1

With our new model in place, we can now describe the two experiments we performed to test our hypotheses about the reasons for the Mr. Chips model's performance being below that of humans in terms of average saccade length and word skipping rates. In Experiment 1, we test the hypothesis that one of the reasons that its performance was below humans is due to its assumption that the goal of the reader is to identify each word with 100% confidence. Specifically, we compare the performance of our model using a 100% criterion vs. a 90% criterion. The former is equivalent to Mr. Chips except for the more realistic motor error function, so for ease of exposition, we will refer to it simply as Mr. Chips.

2.4.1 Methods

Confidence criterion

We set $\alpha = 1.0$ to replicate Mr. Chips, and $\alpha = 0.9$ for the model using a slightly lower confidence criterion to trigger moving on to the next word.

Language model

Both models used a unigram language model, smoothed with Kneser-Ney under default parameters (Chen & Goodman, 1998; equivalent to add-δ smoothing for a unigram model). As in Legge et al. (2002), the models were trained on a 280,000 word corpus of *Grimms' Fairy Tales*, containing 7503 unique

Table 2.1: Mean saccade size (and std. error) for each model

Model	Mean saccade size
Mr. Chips	6.7 (.012)
Without context, 90% criterion	7.1 (.013)
With context, 90% criterion	7.5 (.014)
Humans	$\approx$ 8 (Rayner, 1998)

words. This corpus was normalized by Legge et al. to convert all letters to lowercase, remove all punctuation other than apostrophes, convert all numbers to their alphabetic equivalents, and remove all gibberish words.

Text

Following Legge et al. (2002), we tested the models by simulating the reading of 40,000 words of text generated from the language model, ensuring that the reading models had a normative probability model for the text they were reading.

2.4.2 Results

The results for mean saccade size for both models are given in the top two rows of Table 2.1. As shown in the table, using a criterion of 90% increases the average size of saccades, bringing it a bit closer to the human estimate of about 8 characters. The results for word skipping rates for the two models are plotted as the lower two lines in Figure 2.1. The results show a modest increase in word skipping rates across almost all word lengths for the new model with a lower criterion, bringing it closer to human performance.

2.4.3 Discussion

Although the gain is modest, the results of Experiment 1 show that changing the goal of the model to one more similar to that of human readers, i.e., relaxing the assumption that words need to be identified with 100% certainty, alters the performance of the model across two measures to look more like human performance. Such a result adds some support to the idea that the relevant

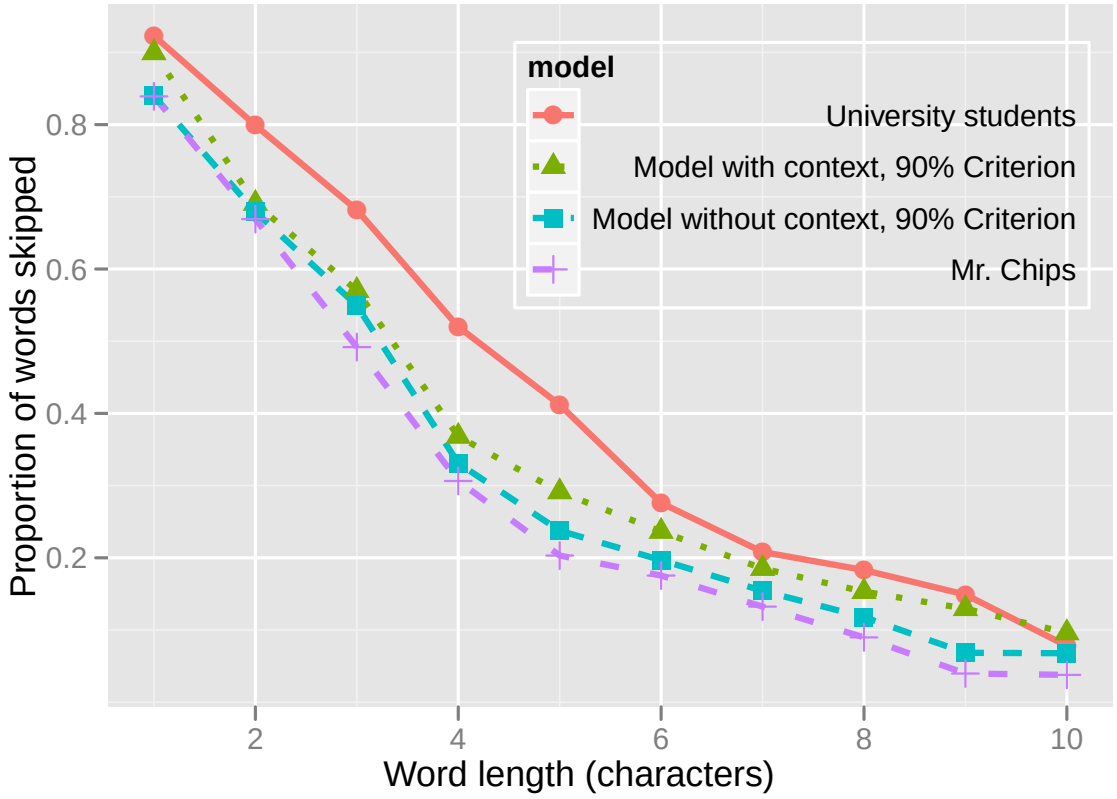

Figure 2.1: Proportion of words skipped by word length for each model. In all cases, the standard error of the mean for the Normal approximation to the Binomial distribution is smaller than the symbols. The human data is from Rayner and McConkie (1976) and has no standard errors.

human behavior is well understood as a rational response to the demands of the task. It is also worth pointing out that the resulting model maintains and uses uncertainty about previous input, something for which most models of sentence processing do not allow.

2.5 Experiment 2

In Experiment 2, we test the effect of allowing the model to use the linguistic context as another source of information for word identification. Specifically, we compare the previous two models to one that includes a 90% confidence criterion as well as a simple bigram model of linguistic context.

2.5.1 Methods

The methods are the same as Experiment 1, except that the new model uses a bigram language model, again smoothed with Kneser-Ney under default parameters.

2.5.2 Results

The results for average saccade length for the new model is given in the third row of Table 2.1. As shown in the table, giving the model information about linguistic structure increases the average size of saccades a bit more, bringing it still closer to the human estimate of 8 characters.

The results for word skipping rates for the new model is plotted as the second line in Figure 2.1. This new model gives an even larger increase in word skipping rates across all word lengths, on top of the increase seen previously, bringing it more in line with human results. Skipping rates are 30% closer to humans than the previous 90% criterion model on average, and for some word lengths are up to 75% closer.

2.5.3 Discussion

The results of this experiment show a case in which making more of the information that is available to a human reader also available to a rational model causes its behavior to more closely approximate human performance. Together with the previous result, this supports the notion that some aspects of reading are well understood as a rational response to the structure of the problem.

2.6 General Discussion

In this paper, we presented a new rational model of reading based on Mr. Chips, but which fixes two problems with it – it uses information about linguistic context in word identification and a flexible identification criterion. Experiment 1 showed that the model's performance looks more like humans' when

the model's goal is shifted to one more like that of humans, 90% confidence in each word. Experiment 2 showed the model's behavior looks even more like humans' when the model can use information that is used by humans: linguistic context. Taken together, these results suggest that many facets of human reading behavior can be well explained as resulting from a rational solution to the problem of reading. This model adds to the growing literature on rational process models, exploring the extent to which human performance can be viewed as rational agents across a wide variety of complex behaviors, such as multiple object tracking (Vul, Frank, Alvarez, & Tenenbaum, 2009) and online change detection (Brown & Steyvers, 2009).

It is the case, however, that many aspects of human reading behavior cannot in principle be explained by a model such as those described in this paper. This is because much of what we know about human reading behavior is about fixation durations, and these models have no notion of duration. They cannot have a notion of duration because visual input is veridical categorical information about a range of characters, so that there is no reason to stay at a given location for more than one timestep. Reichle and Laurent (2006) overcome this problem by making the simplifying assumption that required processing times on words are a function only of their length.

We believe, however, that the way forward for rational models of reading is to incorporate a model of noisy visual input, so that the model can make decisions about fixation durations to get more or less visual information. In other work (Bicknell & Levy, 2010b), we explore the use of such models to answer questions that are impossible to ask with non-rational models of reading such as when and why should between-word regressions be made, and how should reading behavior change as accuracy is valued more or less relative to speed.

2.7 Acknowledgements

Chapter 2, in full, is an exact copy of the material as it appears in Bicknell and Levy (2010a) [Rational eye movements in reading combining uncertainty about previous words and contextual probability. In S. Ohlsson & R. Catram-

bone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1142–1147). Austin, TX: Cognitive Science Society.] 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 84th Annual Meeting of the Linguistic Society of America.

Chapter 3

A rational model of eye movement control in reading

Klinton Bicknell and Roger Levy

Abstract. A number of results in the study of real-time sentence comprehension have been explained by computational models as resulting from the rational use of probabilistic linguistic information. Many times, these hypotheses have been tested in reading by linking predictions about relative word difficulty to word-aggregated eye tracking measures such as go-past time. In this paper, we extend these results by asking to what extent reading is well-modeled as rational behavior at a finer level of analysis, predicting not aggregate measures, but the duration and location of each fixation. We present a new rational model of eye movement control in reading, the central assumption of which is that eye movement decisions are made to obtain noisy visual information as the reader performs Bayesian inference on the identities of the words in the sentence. As a case study, we present two simulations demonstrating that the model gives a rational explanation for between-word regressions.

3.1 Introduction

The language processing tasks of reading, listening, and even speaking are remarkably difficult. Good performance at each one requires integrating a range of types of probabilistic information and making incremental predictions on the basis of noisy, incomplete input. Despite these requirements, empirical work has shown that humans perform very well (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Sophisticated models have been developed that explain many of these effects using the tools of computational linguistics and large-scale corpora to make normative predictions for optimal performance in these tasks (Genzel & Charniak, 2002, 2003; Keller, 2004; Levy & Jaeger, 2007; Jaeger, 2010). To the extent that the behavior of these models looks like human behavior, it suggests that humans are making rational use of all the information available to them in language processing. In the domain of incremental language comprehension, especially, there is a substantial amount of computational work suggesting that humans behave rationally (e.g., Jurafsky, 1996; Narayanan & Jurafsky, 2001; Levy, 2008; Levy, Reali, & Griffiths, 2009). Most of this work has taken as its task predicting the difficulty of each word in a sentence, a major result being that a large component of the difficulty of a word appears to be a function of its probability in context (Hale, 2001; Smith & Levy, 2008). Much of the empirical basis for this work comes from studying reading, where word difficulty can be related to the amount of time that a reader spends on a particular word. To relate these predictions about word difficulty to the data obtained in eye tracking experiments, the eye movement record has been summarized through word aggregate measures, such as the average duration of the first fixation on a word, or the amount of time between when a word is first fixated and when the eyes move to its right ('go-past time').

It is important to note that this notion of word difficulty is an abstraction over the actual task of reading, which is made up of more fine-grained decisions about how long to leave the eyes in their current position, and where to move them next, producing the series of relatively stable periods (fixations) and movements (saccades) that characterize the eye tracking record. While there

has been much empirical work on reading at this fine-grained scale (see Rayner, 1998 for an overview), and there are a number of successful models (Reichle et al., 2006; Engbert et al., 2005), little is known about the extent to which human reading behavior appears to be rational at this finer grained scale. In this paper, we present a new rational model of eye movement control in reading, the central assumption of which is that eye movement decisions are made to obtain noisy visual information, which the reader uses in Bayesian inference about the form and structure of the sentence. As a case study, we show that this model gives a rational explanation for between-word regressions.

In Section 3.2, we briefly describe the leading models of eye movements in reading, and in Section 3.3, we describe how these models account for between-word regressions and the intuition behind our model's account of them. Section 3.4 describes the model and its implementation and Sections 3.5–3.6 describe two simulations we performed with the model comparing behavioral policies that make regressions to those that do not. In Simulation 1, we show that specific regressive policies outperform specific non-regressive policies, and in Simulation 2, we use optimization to directly find optimal policies for three performance measures. The results show that the regressive policies outperform non-regressive policies across a wide range of performance measures, demonstrating that our model predicts that making between-word regressions is a rational strategy for reading.

3.2 Models of eye movements in reading

The two most successful models of eye movements in reading are *E-Z* Reader (Reichle et al., 1998, 2006) and SWIFT (Engbert et al., 2002, 2005). Both of these models characterize the problem of reading as one of word identification. In *E-Z* Reader, for example, the system identifies each word in the sentence serially, moving attention to the next word in the sentence only after processing the current word is complete, and (to slightly oversimplify), the eyes then follow the attentional shifts at some lag. SWIFT works similarly, but with the main difference being that processing and attention are distributed over

multiple words, such that adjacent words can be identified in parallel. While both of these models provide a good fit to eye tracking data from reading, neither model asks the higher level question of what a rational solution to the problem would look like.

The first model to ask this question, Mr. Chips (Legge et al., 1997, 2002), predicts the optimal sequence of saccade targets to read a text based on a principle of minimizing the expected entropy in the distribution over identities of the current word. Unfortunately, however, the Mr. Chips model simplifies the problem of reading in a number of ways: First, it uses a unigram model as its language model, and thus fails to use any information in the linguistic context to help with word identification. Second, it only moves on to the next word after unambiguous identification of the current word, whereas there is experimental evidence that comprehenders maintain some uncertainty about the word identities. In other work, we have extended the Mr. Chips model to remove these two limitations, and show that the resulting model more closely matches human performance (Bicknell & Levy, 2010a). The larger problem, however, is that each of these models uses an unrealistic model of visual input, which obtains absolute knowledge of the characters in its visual window. Thus, there is no reason for the model to spend longer on one fixation than another, and the model only makes predictions for *where* saccades are targeted, and not *how long* fixations last.

Reichle and Laurent (2006) presented a rational model that overcame the limitations of Mr. Chips to produce predictions for both fixation durations and locations, focusing on the ways in which eye movement behavior is an adaptive response to the particular constraints of the task of reading. Given this focus, Reichle and Laurent used a very simple word identification function, for which the time required to identify a word was a function only of its length and the relative position of the eyes. In this paper, we present another rational model of eye movement control in reading that, like Reichle and Laurent, makes predictions for fixation durations and locations, but which focuses instead on the dynamics of word identification at the core of the task of reading. Specifically, our model identifies the words in a sentence by performing Bayesian inference

combining noisy input from a realistic visual model with a language model that takes context into account.

3.3 Explaining between-word regressions

In this paper, we use our model to provide a novel explanation for between-word regressive saccades. In reading, about 10–15% of saccades are regressive – movements from right-to-left (or to previous lines). To understand how models such as *E-Z* Reader or SWIFT account for regressive saccades to previous words, recall that the system identifies words in the sentence (generally) left to right, and that identification of a word in these models takes a certain amount of time and then is completed. In such a setup, why should the eyes ever move backwards? Three major answers have been put forward. One possibility given by *E-Z* Reader is as a response to overshoot; i.e., the eyes move backwards to a previous word because they accidentally landed further forward than intended due to motor error. Such an explanation could only account for small between-word regressions, of about the magnitude of motor error. The most recent version, *E-Z* Reader 10 (Reichle et al., 2009), has a new component that can produce longer between-word regressions. Specifically, the model includes a flag for postlexical integration failure, that – when triggered – will instruct the model to produce a between-word regression to the site of the failure. That is, between-word regressions in *E-Z* Reader 10 can arise because of postlexical processes external to the model's main task of word identification. A final explanation for between-word regressions, which arises as a result of normal processes of word identification, comes from the SWIFT model. In the SWIFT model, the reader can fail to identify a word but move past it and continue reading. In these cases, there is a chance that the eyes will at some point move back to this unidentified word to identify it. From the present perspective, however, it is unclear how it could be rational to move past an unidentified word and decide to revisit it only much later.

Here, we suggest a new explanation for between-word regressions that arises as a result of word identification processes (unlike that of *E-Z* Reader)

and can be understood as rational (unlike that of SWIFT). Whereas in SWIFT and *E-Z* Reader, word recognition is a process that takes some amount of time and is then 'completed', some experimental evidence suggests that word recognition may be best thought of as a process that is never 'completed', as comprehenders appear to both maintain uncertainty about the identity of previous input and to update that uncertainty as more information is gained about the rest of the sentence (Connine, Blasko, & Hall, 1991; Levy, Bicknell, et al., 2009). Thus, it is possible that later parts of a sentence can cause a reader's confidence in the identity of the previous regions to fall. In these cases, a rational way to respond might be to make a between-word regressive saccade to get more visual information about the (now) low confidence previous region.

To illustrate this idea, consider the case of a language composed of just two strings, *AB* and *BA*, and assume that the eyes can only get noisy information about the identity of one character at a time. After obtaining a little information about the identity of the first character, the reader may be reasonably confident that its identity is *A* and move on to obtaining visual input about the second character. If the first noisy input about the second character also indicates that it is probably *A*, then the normative probability that the first character is *A* (and thus a rational reader's confidence in its identity) will fall. This simple example just illustrates the point that if a reader is combining noisy visual information with a language model, then confidence in previous regions will sometimes fall.

There are two ways that a rational agent might deal with this problem. The first option would be to reach a higher level of confidence in the identity of each word before moving on to the right, i.e., slowing down reading left-to-right to prevent having to make right-to-left regressions. The second option is to read left-to-right relatively more quickly, and then make occasional right-to-left regressions in the cases where probability in previous regions falls. In this paper, we present two simulations suggesting that when using a rational model to read natural language, the best strategies for coping with the problem of confidence about previous regions dropping – for any trade-off between speed and accuracy – involve making between-word regressions. In the next section, we present the details of our model of reading and its implementation, and then

we present our two simulations in the sections following.

3.4 Reading as Bayesian inference

At its core, the framework we are proposing is one of reading as Bayesian inference. Specifically, the model begins reading with a prior distribution over possible identities of a sentence given by its language model. On the basis of that distribution, the model decides whether or not to move its eyes (and if so where to move them to) and obtains noisy visual input about the sentence at the eyes' position. That noisy visual input then gives the likelihood term in a Bayesian belief update, where the model's prior distribution over the identity of the sentence given the language model is updated to a posterior distribution taking into account both the language model and the visual input obtained thus far. On the basis of that new distribution, the model again selects an action and the cycle repeats.

This framework is unique among models of eye movement control in reading (except Mr. Chips) in having a fully explicit model of how visual input is used to discriminate word identity. This approach stands in sharp contrast to other models, which treat the time course of word identification as an exogenous function of other influencing factors (such as word length, frequency, and predictability). The hope in our approach is that the influence of these key factors on the eye movement record will fall out as a natural consequence of rational behavior itself. For example, it is well known that the higher the conditional probability of a word given preceding material, the more rapidly that word is read (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008; Ehrlich & Rayner, 1981; Smith & Levy, 2008). *E-Z* Reader and SWIFT incorporate this finding by specifying a dependency on word predictability in the exogenous function determining word processing time. In our framework, in contrast, we would expect such an effect to emerge as a byproduct of Bayesian inference: words with high prior probability (conditional on preceding fixations) will require less visual input to be reliably identified.

An implemented model in this framework must formalize a number of

pieces of the reading problem, including the possible actions available to the reader and their consequences, the nature of visual input, a means of combining visual input with prior expectations about sentence form and structure, and a control policy determining how the model will choose actions on the basis of its posterior distribution over the identities of the sentence. In the remainder of this section, we present these details of the formalization of the reading problem we used for the simulations reported in this paper: actions (3.4.1), visual input (3.4.2), formalization of the Bayesian inference problem (3.4.3), control policy (3.4.4), and finally, implementation of the model using weighted finite state automata (3.4.5).

3.4.1 Formal problem of reading: Actions

For our model, we assume a series of discrete timesteps, and on each time step, the model first obtains visual input around the current location of the eyes, and then chooses between three actions: (a) continuing to fixate the currently fixated position, (b) initiating a saccade to a new position, or (c) stopping reading of the sentence. If on the ith timestep, the model chooses option (a), the timestep advances to $i + 1$ and another sample of visual input is obtained around the current position. If the model chooses option (c), the reading immediately ends. If a saccade is initiated (b), there is a lag of two timesteps, roughly representing the time required to plan and execute a saccade, during which the model again obtains visual input around the current position and then the eyes move – with some motor error – toward the intended target t_i, landing on position ℓ_i. On the next time step, visual input is obtained around ℓ_i and another decision is made. The motor error for saccades follows the form of random error used by all major models of eye movements in reading: the landing position ℓ_i is normally distributed around the intended target t_i with standard deviation

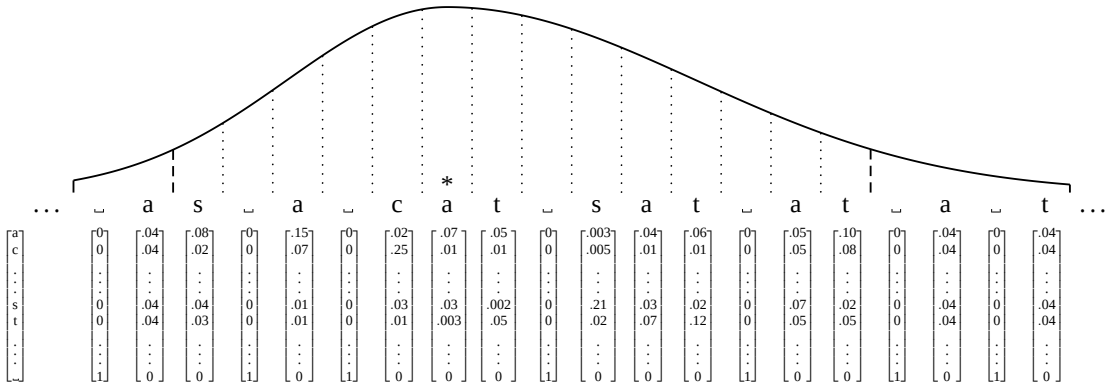

Figure 3.1: Peripheral and foveal visual input in the model. The asymmetric Gaussian curve indicates declining perceptual acuity centered around the fixation point (marked by $*$). The vector under each letter position denotes the likelihood $p(\mathcal{I}(j)|w_j)$ for each possible letter w_j, taken from a single input sample with $\Lambda = 1/\sqrt{3}$ (see vector at the left edge of the figure for key, and Section 3.4.2). In peripheral vision, the distinction of letters from whitespace is veridical, but no information about letter identity is obtained. Note in this particular sample, input from the fixated character and the following one is rather inaccurate.

given by a linear function of the intended distance[1]

$$\ell_i \sim \mathcal{N}\left(t_i, (\delta_0 + \delta_1|t_i - \ell_{i-1}|)^2\right) \tag{3.1}$$

for some linear coefficients δ_0 and δ_1. In the experiments reported in this paper, we follow the SWIFT model in using $\delta_0 = 0.87, \delta_1 = 0.084$.

3.4.2 Noisy visual input

As stated earlier, the role of noisy visual input in our model is as the likelihood term in a Bayesian inference about sentence form and identity. Therefore, if we denote the input obtained thus far from a sentence as $\mathcal{I}$, all the information pertinent to the reader's inferences can be encapsulated in the form $p(\mathcal{I}|w)$ for possible sentences w. We assume that the inputs deriving from each character position are conditionally independent given sentence identity, so that if w_j denotes letter j of the sentence and $\mathcal{I}(j)$ denotes the component of visual input

[1] In the terminology of the literature, the model has only random motor error (variance), not systematic error (bias). Following Engbert and Krügel (2010), systematic error may arise from Bayesian estimation of the best saccade distance.

associated with that letter, then we can decompose $p(\mathcal{I}|w)$ as $\prod_j p(\mathcal{I}(j)|w_j)$. For simplicity, we assume that each character is either a lowercase letter or a space. The visual input obtained from an individual fixation can thus be summarized as a vector of likelihoods $p(\mathcal{I}(j)|w_j)$, as shown in Figure 3.1. As in the real visual system, our visual acuity function decreases with retinal eccentricity; we follow the SWIFT model in assuming that the spatial distribution of visual processing rate follows an asymmetric Gaussian with $\sigma_L = 2.41, \sigma_R = 3.74$, which we discretize into processing rates for each character position. If ϵ denotes a character's eccentricity in characters from the center of fixation, then the proportion of the total processing rate at that eccentricity $\lambda(\epsilon)$ is given by integrating the asymmetric Gaussian over a character width centered on that position,

$$\lambda(\epsilon) = \int_{\epsilon-.5}^{\epsilon+.5} \frac{1}{Z} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx, \sigma = \begin{cases} \sigma_L, & x < 0 \\ \sigma_R, & x \geq 0 \end{cases}$$

where the normalization constant Z is given by

$$Z = \sqrt{\frac{\pi}{2}}(\sigma_L + \sigma_R).$$

From this distribution, we derive two types of visual input, *peripheral* input giving word boundary information and *foveal* input giving information about letter identity.

Peripheral visual input

In our model, any eccentricity with a processing rate proportion $\lambda(\epsilon)$ at least 0.5% of the rate proportion for the centrally fixated character ($\epsilon \in [-7, 12]$), yields peripheral visual input, defined as veridical word boundary information indicating whether each character is a letter or a space. This roughly corresponds to empirical estimates that humans obtain useful information in reading from about 19 characters, more from the right of fixation than the left (Rayner, 1998). Hence in Figure 3.1, for example, left-peripheral visual input can be represented as veridical knowledge of the initial whitespace (denoted ␣), and a uniform distribution over the 26 letters of English for the letter a.

Foveal visual input

In addition, for those eccentricities with a processing rate proportion $\lambda(\epsilon)$ that is at least 1% of the total processing rate ($\epsilon \in [-5, 8]$) the model receives foveal visual input, defined only for letters[2] to give noisy information about the letter's identity. This threshold of 1% roughly corresponds to estimates that readers get information useful for letter identification from about 4 characters to the left and 8 to the right of fixation (Rayner, 1998).

In our model, each letter is equally confusable with all others, following Norris (2006, 2009), but ignoring work on letter confusability (which could be added to future model revisions; Engel, Dougherty, & Brian Jones, 1973; Geyer, 1977). Visual information about each character is obtained by sampling. Specifically, we represent each letter as a 26-dimensional vector, where a single element is 1 and the other 25 are zeros, and given this representation, foveal input for a letter is given as a sample from a 26-dimensional Gaussian with a mean equal to the letter's true identity and a diagonal covariance matrix $\Sigma(\epsilon) = \lambda(\epsilon)^{-1/2} I$. It is relatively straightforward to show that under these conditions, if we take the processing rate to be the expected change in log-odds of the true letter identity relative to any other that a single sample brings about, then the rate equals $\lambda(\epsilon)$. We scale the overall processing rate by multiplying each rate by Λ. For the experiments in this paper, we set $\Lambda = 4$. For each fixation, we sample independently from the appropriate distribution for each character position and then compute the likelihood given each possible letter, as illustrated in the non-peripheral region of Figure 3.1.

3.4.3 Inference about sentence identity

Given the visual input and a language model, inferences about the identity of the sentence w can be made by standard Bayesian inference, where the prior is given by the language model and the likelihood is a function of the total

[2]For white space, the model is already certain of the identity because of peripheral input.

visual input obtained from the first to the ith timestep $\mathcal{I}_1^i$,

$$p(w|\mathcal{I}_1^i) = \frac{p(w)p(\mathcal{I}_1^i|w)}{\sum_{w'}(w')p(\mathcal{I}_1^i|w')}. \tag{3.2}$$

If we let $\mathcal{I}(j)$ denote the input received about character position j and let w_j denote the jth character in sentence identity w, then the likelihood can be broken down by character position as

$$p(\mathcal{I}_1^i|w) = \prod_{j=1}^{n} p(\mathcal{I}_1^i(j)|w_j)$$

where n is the final character about which there is any visual input. Similarly, we can decompose this into the product of the likelihoods of each sample

$$p(\mathcal{I}_1^i|w) = \prod_{j=1}^{n}\prod_{t=1}^{i} p(\mathcal{I}_t(j)|w_j). \tag{3.3}$$

If the eccentricity of the jth character on the tth timestep ϵ_t^j is outside of foveal input or the character is a space, the inner term is 0 or 1. If the sample was from a letter in foveal input $\epsilon_t^j \in [-5, 8]$, it is the probability of sampling $\mathcal{I}_t(j)$ from the multivariate Gaussian $\mathcal{N}(w_j, \Lambda\Sigma(\epsilon_t^j))$.

3.4.4 Control policy

The model uses a simple policy to decide between actions based on the marginal probability m of the most likely character c in position j,

$$\begin{aligned} m(j) &= \max_c p(w_n = c|\mathcal{I}_1^i) \\ &= \max_c \sum_{w':w'_n=c} p(w'|\mathcal{I}_1^i). \end{aligned} \tag{3.4}$$

Intuitively, a high value of m means that the model is relatively confident about the character's identity, and a low value that it is relatively uncertain.

Given the values of this statistic, our model decides between four possible actions, as illustrated in Figure 3.2. If the value of this statistic for the current position of the eyes $m(\ell_i)$ is less than a parameter α, the model chooses to continue fixating the current position (3.2a). Otherwise, if the value of $m(j)$ is less

(a) $m = [.6, .7, \mathbf{.6}, .4, .3, .6]$: Keep fixating (3)
(b) $m = [.6, .4, \mathbf{.9}, .4, .3, .6]$: Move back (to 2)
(c) $m = [.6, .7, \mathbf{.9}, .4, .3, .6]$: Move forward (to 6)
(d) $m = [.6, .7, \mathbf{.9}, .8, .7, .7]$: Stop reading

Figure 3.2: Values of m for a 6 character sentence under which a model fixating position 3 would take each of its four actions, if $\alpha = .7$ and $\beta = .5$.

than β for some leftward position $j < \ell_i$, the model initiates a saccade to the closest such position (3.2b). If $m(j) \geq \beta$ for all $j < \ell_i$, then the model initiates a saccade to n characters past the closest position to the right $j > \ell_i$ for which $m(j) < \alpha$ (3.2c).[3] Finally, if no such positions exist to the right, the model stops reading the sentence (3.2d). Intuitively, then, the model reads by making a rightward sweep to bring its confidence in each character up to α, but pauses to move left if confidence in a previous character falls below β.

3.4.5 Implementation with wFSAs

This model can be efficiently and simply implemented using weighted finite-state automata (wFSAs; Mohri, 1997) as follows: First, we begin with a wFSA representation of the language model, where each arc emits a single character (or is an epsilon-transition emitting nothing). To perform belief update given a new visual input, we create a new wFSA to represent the likelihood of each character from the sample. Specifically, this wFSA has only a single chain of states, where, e.g., the first and second state in the chain are connected by 27 (or fewer) arcs, which emit each of the possible characters for w_1 along with their respective likelihoods given the visual input (as in the inner term of Equation 3.3). Next, these two wFSAs may simply be composed and then normalized, which completes the belief update, resulting in a new wFSA giving the posterior distribution over sentences. To calculate the statistic m, while it is possible to calculate it in closed form from such a wFSA relatively straightforwardly, for efficiency we use Monte Carlo estimation based on samples from the wFSA.

[3]The role of n is to ensure that the model does not center its visual field on the first uncertain character. We did not attempt to optimize this parameter, but fixed n at 2.

3.5 Simulation 1

With the description of our model in place, we next proceed to describe the first simulation in which we used the model to test the hypothesis that making regressions is a rational way to cope with confidence in previous regions falling. Because there is in general no single rational trade-off between speed and accuracy, our hypothesis is that, for any given level of speed and accuracy achieved by a non-regressive policy, there is a faster and more accurate policy that makes a faster left-to-right pass but occasionally does make regressions. In the terms of our model's policy parameters α and β described above, non-regressive policies are exactly those with $\beta = 0$, and a policy that is faster on the left-to-right pass but does make regressions is one with a lower value of α but a non-zero β. Thus, we tested the performance of our model on the reading of a corpus of text typical of that used in reading experiments at a range of reasonable non-regressive policies, as well as a set of regressive policies with lower α and positive β. Our prediction is that the former set will be strictly dominated in terms of both speed and accuracy by the latter.

3.5.1 Methods

Policy parameters

We test 4 non-regressive policies (i.e., those with $\beta = 0$) with values of $\alpha \in \{.90, .95, .97, .99\}$, and in addition, test regressive policies with a lower range of $\alpha \in \{.85, .90, .95, .97\}$ and $\beta \in \{.4, .7\}$.[4]

Language model

Our reader's language model was an unsmoothed bigram model created using a vocabulary set consisting of the 500 most frequent words in the British National Corpus (BNC) as well as all the words in our test corpus. From this

[4]We tested all combinations of these values of α and β except for $[\alpha, \beta] = [.97, .4]$, because we did not believe that a value of β so low in relation to α would be very different from a non-regressive policy.

vocabulary, we constructed a bigram model using the counts from every bigram in the BNC for which both words were in vocabulary (about 222,000 bigrams).

wFSA implementation

We implemented our model with wFSAs using the OpenFST library (Allauzen, Riley, Schalkwyk, Skut, & Mohri, 2007). Specifically, we constructed the model's initial belief state (i.e., the distribution over sentences given by its language model) by directly translating the bigram model into a wFSA in the log semiring. We then composed this wFSA with a weighted finite-state transducer (wFST) breaking words down into characters. This was done in order to facilitate simple composition with the visual likelihood wFSA defined over characters. In the Monte Carlo estimation of m, we used 5000 samples from the wFSA. Finally, to speed performance, we bounded the wFSA to have exactly the number of characters present in the actual sentence and then re-normalized.

Test corpus

We tested our model's performance by simulating reading of the Schilling corpus (Schilling, Rayner, & Chumbley, 1998). To ensure that our results did not depend on smoothing, we only tested the model on sentences in which every bigram occurred in the BNC. Unfortunately, only 8 of the 48 sentences in the corpus met this criterion. Thus, we made single-word changes to 25 more of the sentences (mostly changing proper names and rare nouns) to produce a total of 33 sentences to read, for which every bigram did occur in the BNC.

3.5.2 Results and discussion

For each policy we tested, we measured the average number of timesteps it took to read the sentences, as well as the average (natural) log probability of the correct sentence identity under the model's beliefs after reading ended 'Accuracy'. The results are plotted in Figure 3.3. As shown in the graph, for each non-regressive policy (the circles), there is a regressive policy that outperforms

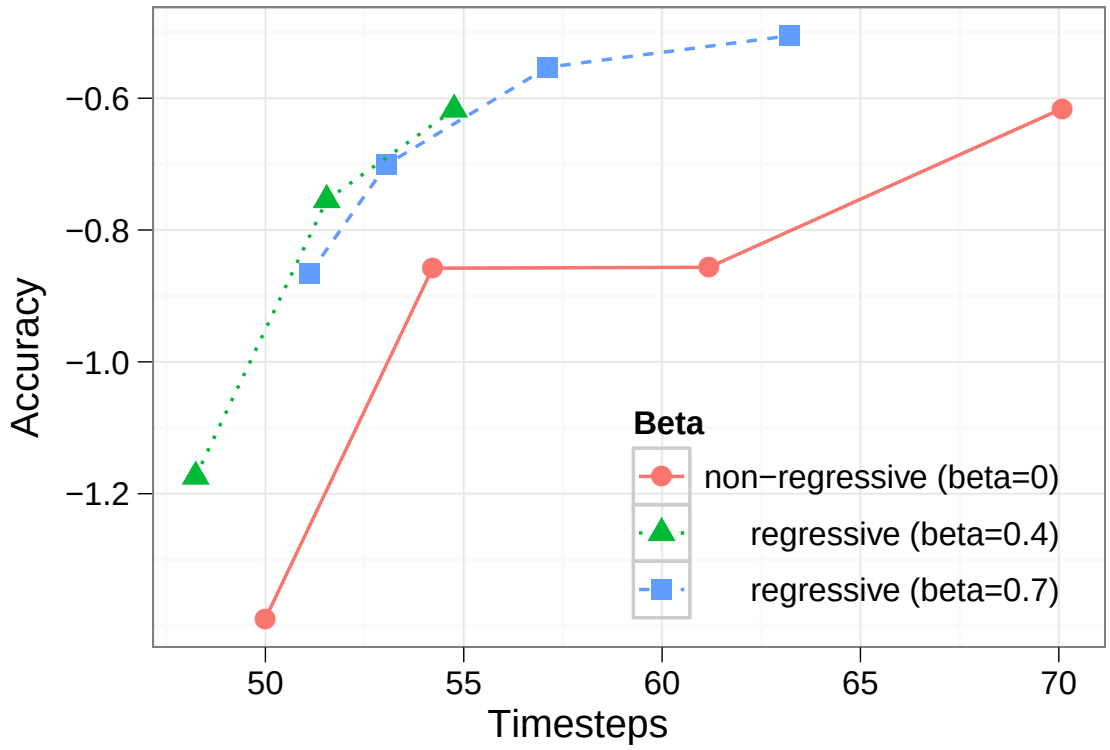

Figure 3.3: Mean number of timesteps taken to read a sentence and (natural) log probability of the true identity of the sentence 'Accuracy' for a range of values of α and β. Values of α are not labeled, but increase with the number of timesteps for a constant value of β. For each non-regressive policy ($\beta = 0$), there is a policy with a lower α and higher β that achieves better accuracy in less time.

it, both in terms of average number of timesteps taken to read (further to the left) and the average log probability of the sentence identity (higher). Thus, for a range of policies, these results suggest that making regressions when confidence about previous regions falls is a rational reader strategy, in that it appears to lead to better performance, both in terms of speed and accuracy.

3.6 Simulation 2

In Simulation 2, we perform a more direct test of the idea that making regressions is a rational response to the problem of confidence falling about previous regions using optimization techniques. Specifically, we search for optimal policy parameter values (α, β) for three different measures of performance,

each representing a different trade-off between the importance of accuracy and speed.

3.6.1 Methods

Performance measures

We examine performance measures interpolating between speed and accuracy of the form

$$L(1-\gamma) - T\gamma \tag{3.5}$$

where L is the log probability of the true identity of the sentence under the model's beliefs at the end of reading, and T is the total number of timesteps before the model decided to stop reading. Thus, each different performance measure is determined by the weighting for time γ. We test three values of $\gamma \in \{.025, .1, .4\}$. The first of these weights accuracy highly, while the final one weights 1 timestep almost as much as 1 unit of log probability.

Optimization of policy parameters

Searching directly for optimal values of α and β for our stochastic reading model is difficult because each evaluation of the model with a particular set of parameters produces a different result. We use the PEGASUS method (Ng & Jordan, 2000) to transform this stochastic optimization problem into a deterministic one on which we can use standard optimization algorithms.[5] Then, we evaluate the model's performance at each value of α and β by reading the full test corpus and averaging performance. We then simply use coordinate ascent (in logit space) to find the optimal values of α and β for each performance measure.

Language model

The language model used in this simulation begins with the same vocabulary set as in Sim. 1, i.e., the 500 most frequent words in the BNC and every

[5] Specifically, this involves fixing the random number generator for each run to produce the same values, resulting in minimizing the variance in performance across evaluations.

Table 3.1: Optimal values of α and β found for each performance measure γ tested and mean performance at those values, measured in timesteps T and (natural) log probability L.

γ	α	β	Timesteps	Log probability
.025	.90	.99	41.2	-0.02
.1	.36	.80	25.8	-0.90
.4	.18	.38	16.4	-4.59

word that occurs in our test corpus. Because the search algorithm demands that we evaluate the performance of our model at a number of parameter values, however, it is too slow to optimize α and β using the full language model that we used for Sim. 1. Instead, we begin with the same set of bigrams used in Sim. 1 – i.e., those that contain two in-vocabulary words – and trim this set by removing rare bigrams that occur less than 200 times in the BNC (except that we do not trim any bigrams that occur in our test corpus). This reduces our set of bigrams to about 19,000.

wFSA implementation

The implementation was the same as in Sim. 1.

Test corpus

The test corpus was the same as in Sim. 1.

3.6.2 Results and discussion

The optimal values of α and β for each $\gamma \in \{.025, .1, .4\}$ are given in Table 3.1 along with the mean values for L and T found at those parameter values. As the table shows, the optimization procedure successfully found values of α and β, which go up (slower reading) as γ goes down (valuing accuracy more than time). In addition, we see that the average results of reading at these parameter values are also as we would expect, with T and L going up as γ goes down. As predicted, the optimal values of β found are non-zero across the range of policies, which include policies that value speed over accuracy much more than in

Sim. 1. This provides more evidence that whatever the particular performance measure used, policies making regressive saccades when confidence in previous regions falls perform better than those that do not.

There is one interesting difference between the results of this simulation and those of Sim. 1, which is that here, the optimal policies all have a value of $\beta > \alpha$. That may at first seem surprising, since the model's policy is to fixate a region until its confidence becomes greater than α and then return if it falls below β. It would seem, then, that the only reasonable values of β are those that are strictly below α. In fact, this is not the case because of the two time step delay between the decision to move the eyes and the execution of that saccade. Because of this delay, the model's confidence when it leaves a region (relevant to β) will generally be higher than when it decided to leave (determined by α). In Simulation 2, because of the smaller grammar that was used, the model's confidence in a region's identity rises more quickly and this difference is exaggerated.

3.7 Conclusion

In this paper, we presented a model that performs Bayesian inference on the identity of a sentence, combining a language model with noisy information about letter identities from a realistic visual input model. On the basis of these inferences, it uses a simple policy to determine how long to continue fixating the current position and where to fixate next, on the basis of information about where the model is uncertain about the sentence's identity. As such, it constitutes a rational model of eye movement control in reading, extending the insights from previous results about rationality in language comprehension.

The results of two simulations using this model support a novel explanation for between-word regressive saccades in reading: that they are used to gather visual input about previous regions when confidence about them falls. Simulation 1 showed that a range of policies making regressions in these cases outperforms a range of non-regressive policies. In Simulation 2, we directly searched for optimal values for the policy parameters for three different performance measures, representing different speed-accuracy trade-offs, and found

that the optimal policies in each case make substantial use of between-word regressions when confidence in previous regions falls. In addition to supporting a novel motivation for between-word regressions, these simulations demonstrate the possibility for testing a range of questions that were impossible with previous models of reading related to the goals of a reader, such as how should reading behavior change as accuracy is valued more.

There are a number of obvious ways for the model to move forward. One natural next step is to make the model more realistic by using letter confusability matrices. In addition, the link to previous work in sentence processing can be made tighter by incorporating syntax-based language models. It also remains to compare this model's predictions to human data more broadly on standard benchmark measures for models of reading. The most important future development, however, will be moving toward richer policy families, which enable more intelligent decisions about eye movement control, based not just on simple confidence statistics calculated independently for each character position, but rather which utilize the rich structure of the model's posterior beliefs about the sentence identity (and of language itself) to make more informed decisions about the best time to move the eyes and the best location to direct them next.

3.8 Acknowledgements

Chapter 3, in full, is an exact copy of the material as it appears in Bicknell and Levy (2010b) [A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL* (pp. 1168–1178). Uppsala, Sweden: Association for Computational Linguistics.] The dissertation author was the primary investigator and author of this paper. In addition to being presented to the Association for Computational Linguistics, this work was also presented at the 23rd Annual CUNY Conference on Human Sentence Processing.

Chapter 4

A rational account of predictability, frequency, and length effects

In the study of eye movements in reading, a range of properties of a text – from low level spatial features to high level discourse structure – have been shown to influence how readers move their eyes. Some of the most robust of these effects have their locus at the level of the word. Specifically, a word's predictability, frequency, and length can have large effects on eye movement behavior. For this reason, current models of eye movement control in reading (e.g., Reichle et al., 1998, 2009; Engbert et al., 2002, 2005) have taken one part of their task to be to reproduce effects of these three linguistic variables, which have been termed part of the 'benchmark phenomena' which 'any viable model of eye-movement control in reading must be able to explain' (Reichle et al., 2003). One of the goals of the present work, then, is to demonstrate that the rational reading framework proposed in the previous chapter[1] can account for these classic effects, which we do by analyzing the behavior produced by simulations run with the model. In addition, since one of the goals of computational modeling is to gain insight into *why* certain effects occur, we will also give intuitions for the reason that the model behaves in this way. Since this is a rational model,

[1]Note that this chapter builds on the previous chapter, and thus presupposes knowledge of much of the material presented there. As a brief reminder of the model parameters that will be relevant for the present chapter, recall that the threshold of confidence in the current character's identity required for the model to initiate a saccade is α, the weighting of time in the reward function is γ, and the weighting of accuracy is $1 - \gamma$.

such an explanation necessarily also constitutes a proposal for why such effects may arise from rational actions of an adaptive agent.

The plan for the rest of this chapter is as follows. The following section gives intuitions for why the model should produce the well-known effects of word predictability, frequency, and length. Section 4.2 describes the results of a simulation demonstrating the effects of predictability, frequency, and length that are produced by our current implementation of the model. Section 4.3 reports the results of another simulation showing the effects of these variables produced by a version of our model that does not make use of linguistic context to help identify words. Finally, in Section 4.4 we summarize the current results and make suggestions for how to remedy some of the model's shortcomings.

4.1 Intuitions

The general findings about the effects of word predictability, frequency, and length on eye movements in reading can be summarized relatively simply: words that are less predictable, lower frequency, and longer tend to receive more and longer fixations (Rayner, 1998, 2009). In this section, we describe intuitions for why our model should reproduce these effects.

4.1.1 Predictability

The basic intuition for why the model should produce effects of word predictability is very closely related to the reason for frequency effects in lexical recognition given by Norris (2006, 2009). This intuition is clearest if we make the simplifying assumption that prior to obtaining any visual information about a word, the model has near-veridical knowledge of the preceding context. In that case, the probability of the true identity of that word in the model is given by the word's predictability in context, which we will denote by π. Now, imagine that the reader begins to fixate the word and gather visual input. On average, this visual input will increase the probability of the true identity of the word under the model's beliefs. The rate of this increase depends on the quality of

visual input – and is largely independent of π.[2] Recall that under our behavior policy, the eyes will remain in this position until the model's confidence in the identity of the character at that position exceeds the threshold α. Because information is being obtained about the entire word simultaneously, the probability of the identity of that single character is closely tied to the identity of the entire word. Thus, the initial probability of the true identity of the fixated character is also likely to start near the initial probability π of the true word, and the model's confidence about the identity of the fixated character is likely to reach the threshold α near the same time that the model's confidence about the identity of the fixated word reaches the threshold. As a consequence, the amount of visual input that is needed to reach the threshold which initiates a saccade is largely a function of the distance between π and α. For more predictable words, π is closer to α, and thus less visual information will be needed on average to reach α. Because less visual input is required, this translates into shorter fixation durations, more skipping, and fewer refixations.

4.1.2 Frequency

The most obvious intuition for the effect of frequency in the model is parasitic on the effect of predictability: words that are lower frequency are likely to be less predictable on average. Thus, the same predictions made for words of higher predictability – shorter fixation durations, more skipping, fewer refixations – should also hold on average for words of high frequency. Unlike in models such as *E-Z* Reader and SWIFT, there is no effect of frequency independent of predictability built into the model.

4.1.3 Length

The predictions for the effect of word length in the model are less clear, as there are at least five ways in which word length might affect eye movements produced by the model. Three of these effects make predictions in line with the

[2] Technically, it is only the rate of the increase in log-odds space that is largely independent of π.

empirical data. First, because of the inverse correlation between word length and frequency, shorter words are more likely to be higher frequency, and thus higher predictability.[3] This would predict correctly that shorter words exhibit shorter fixation durations, less skipping, and fewer refixations. Second, less visual input about the end of longer words can be obtained parafoveally, which could motivate a similar effect. Third, because of the motor error in the model, longer words are more likely to be unintentionally fixated or refixated, which could contribute to the same pattern of skipping and refixation effects.

However, there are also at least two reasons which might lead shorter words to have longer fixation durations and more skipping refixations, both of which arise from limitations of the current implementation of the model, and both of which relate to the relationship between word length and neighborhood size. Neighborhood size should have a strong effect on any model of word identification from noisy visual input (e.g., Norris, 2006, 2009), since a word that has many likely neighbors will require more visual input for a reader to be confident in its identity than a word with fewer neighbors. Because we removed a substantial number of lower frequency words from the grammar in order to speed up the current simulations,[4] longer words will have artificially few neighbors, which could in turn speed their recognition. Additionally, recall that the model has veridical knowledge of word boundaries 6 characters to the left and 14 characters to the right of the point of fixation, which means that in almost all cases, it will have veridical knowledge of the exact length of the word it is fixating. For humans, however, it seems reasonable to assume that the absolute amount of uncertainty about a word's exact length is larger for longer words; e.g., the chance that a reader would mistake a word that was actually 2 letters long for being 3 letters seems intuitively smaller than the chance of mistaking a word that was actually 12 letters for being 13.[5] Under this assumption, then

[3] In fact, Piantadosi, Tily, and Gibson (2011) have shown that word length is even more closely correlated with average predictability than with frequency.

[4] For details of this procedure, see the previous chapter.

[5] There could be a number of possible forms this relationship may take: perhaps the variance in people's estimates of word length is constant in log space, or perhaps it is simply proportional to the length. All that is required for the argument presented here is that the absolute amount of uncertainty increases for longer words.

the effect of the model's veridical knowledge will also be to artificially speed recognition especially for longer words. In sum, while there are three reasons to expect the model to produce length effects in the same direction as human readers, there are at least two technical limitations of the current implementation of the model that might be expected to produce length effects in the opposite direction.

4.2 Simulation 1: full model

Having described the intuitions behind why the model might be expected to produce effects of predictability, frequency, and length, we now assess the effects of these variables that the model does in fact produce. To do so, we used the model to simulate reading of the entirety of our modified version of the Schilling corpus 100 times.

For all three variables, successful model behavior is taken largely to be producing monotonic effects in the same direction as those generally found for humans. In addition, effects of one of these three variables, frequency, may be judged according to a higher standard, as frequency effects on aggregate measures have been well documented for human readers of the Schilling corpus. For example, Pollatsek et al. (2006) reports the mean human values of these four aggregate measures for each of five frequency classes. Thus, for frequency, we can also directly compare the predictions of the model to human reading behavior.

4.2.1 Methods

The methodology for running simulations with the model is largely identical to that used in Simulation 2 of the previous chapter. The only difference is that our reward function was parameterized by a single value, $\gamma = .05$.

For effects of frequency, we can directly compare the model's reading behavior to that of humans. Of the four aggregate measures, this comparison is straightforward to make for skip rates and refixation rates, but to compare fixation durations, we must first convert the model's raw timesteps into millisec-

onds. We performed this scaling by multiplying the duration of each fixation (in timesteps) by a conversion factor set to be equal to the mean human gaze duration (in milliseconds) divided by the mean model gaze duration (in timesteps) for the highest frequency bin. That is, we performed scaling such that the model predictions exactly matched the human data for gaze durations in the highest frequency bin.

4.2.2 Results

From the output of the simulations, we analyzed the effects of word predictability, frequency, and length on four aggregate measures of eye movements in first pass reading (i.e., prior to any fixations beyond the current word): the duration of the first fixation on a word during first pass reading ('first fixation duration'), the total duration of the first unbroken sequence of fixations made on a word during first pass reading ('gaze duration'), the proportion of trials in which a word was not fixated on first pass ('skip rate'), and the proportion of trials in which the first unbroken sequence of fixations made on a word in first pass reading comprised more than one fixation ('refixation rate'). All durations are reported in model timesteps except for the frequency results, which are reported in milliseconds to facilitate comparison with human data.

Predictability

Figure 4.1 shows the effect of predictability on the four aggregate measures, as estimated by loess from means calculated for each word token in the corpus. As predicted by both the intuition given above and empirical human data, there are shorter fixations, more skipping, and fewer refixations for more predictable words.

Frequency

Figure 4.2 shows the effects of frequency (binned by rounding down to the nearest integer, in order to facilitate comparison to the means reported by

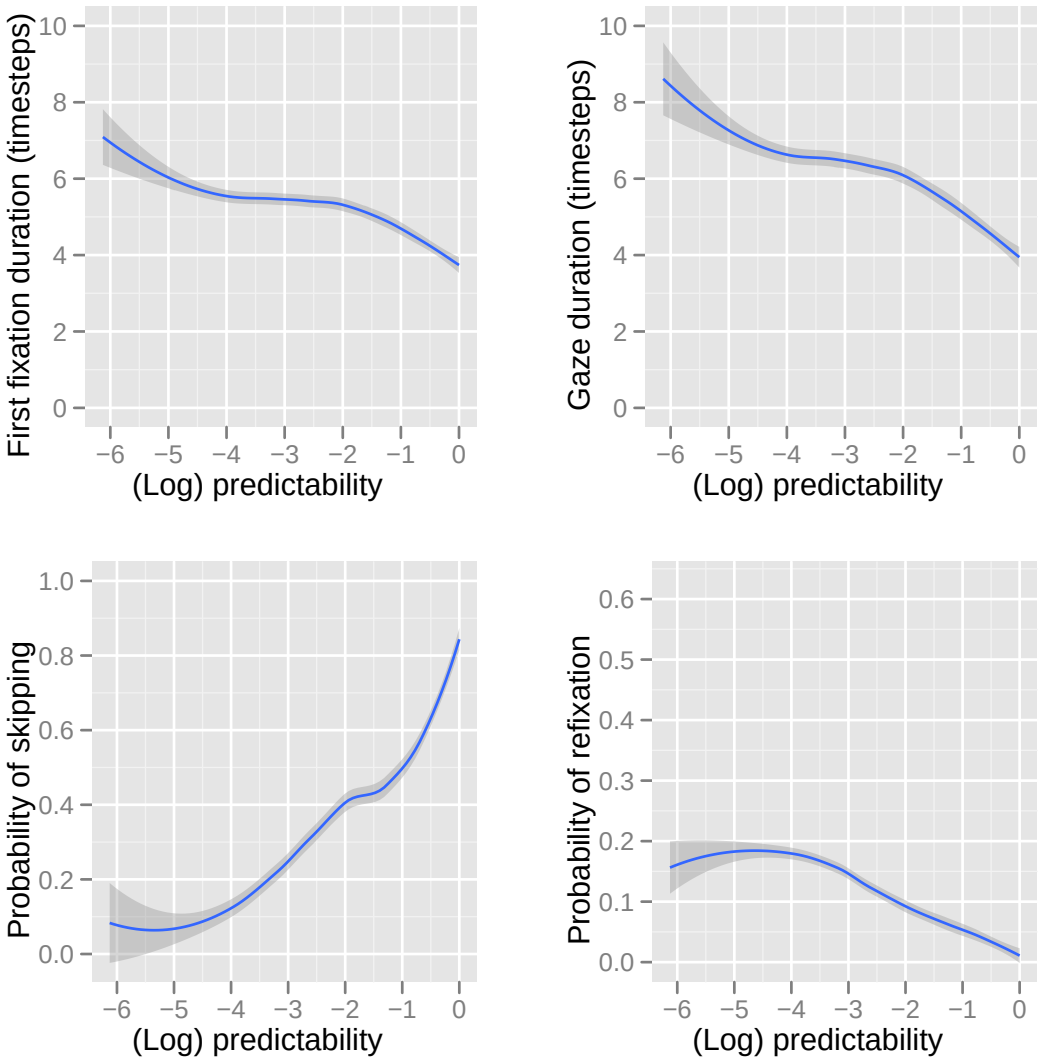

Figure 4.1: The full model's predicted effect of word predictability on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation, as estimated from simulations by loess, plotted with standard errors calculated across word tokens.

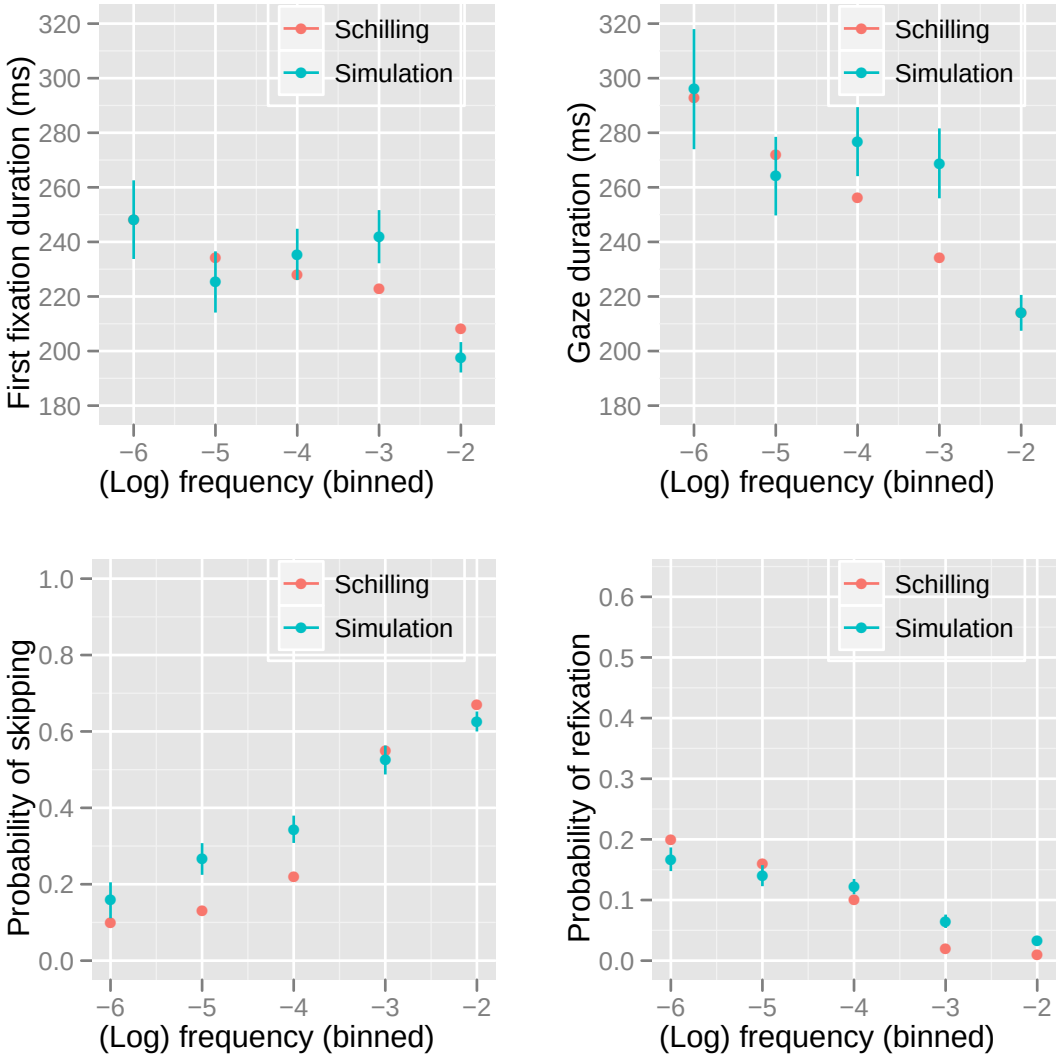

Figure 4.2: The full model's predicted effect of word frequency on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation, as estimated from simulations, plotted with standard errors calculated across word tokens. Mean values from the Schilling corpus reported by Pollatsek et al. (2006) are shown for comparison.

Pollatsek et al., 2006) on the four aggregate measures. Comparing just the highest and lowest frequency bins of the model and the human data shows a striking fit in effect direction and magnitude for all four measures. Further, predictions for skip rates and refixation rates appear to be close fits to the human data throughout the range. The effect of frequency on first fixation durations and gaze durations, however, is a worse fit for intermediate frequencies and does not appear to be completely monotonic.

Length

Figure 4.3 shows the effects of word length on the four aggregate measures, as estimated by loess from means calculated for each word token in the corpus. Inspecting the graphs, we see effects in the same direction as human readers for words of length 1–4: within that range, words that are longer have longer fixations, less skipping, and a higher rate of refixations. For lengths greater than 4, however, we see a different pattern. For this range, there appear to be length effects in the wrong direction for fixation durations, a smaller effect in the correct direction for skipping rates, and no effect for refixation rates.

4.2.3 Discussion

In summary, these results demonstrate that effects of predictability, frequency, and length in the behavior of our full model resemble that of human readers in many aspects. Predictability effects on all four aggregate measures are monotonic and in the same direction as predicted. Frequency effects (which we can compare quantitatively with human readers) on all four measures are in the same direction as predicted, and the total magnitude of the effect (i.e., the difference between the highest and lowest frequency bins) is strikingly similar to that displayed by human readers, despite the fact that we have not made any attempt to fit the human data.[6] In between the extreme frequency bins, skip rates and refixation rates also closely matched human data, but for the duration

[6]With the exception, of course, of the scaling parameter that converts model timesteps to milliseconds.

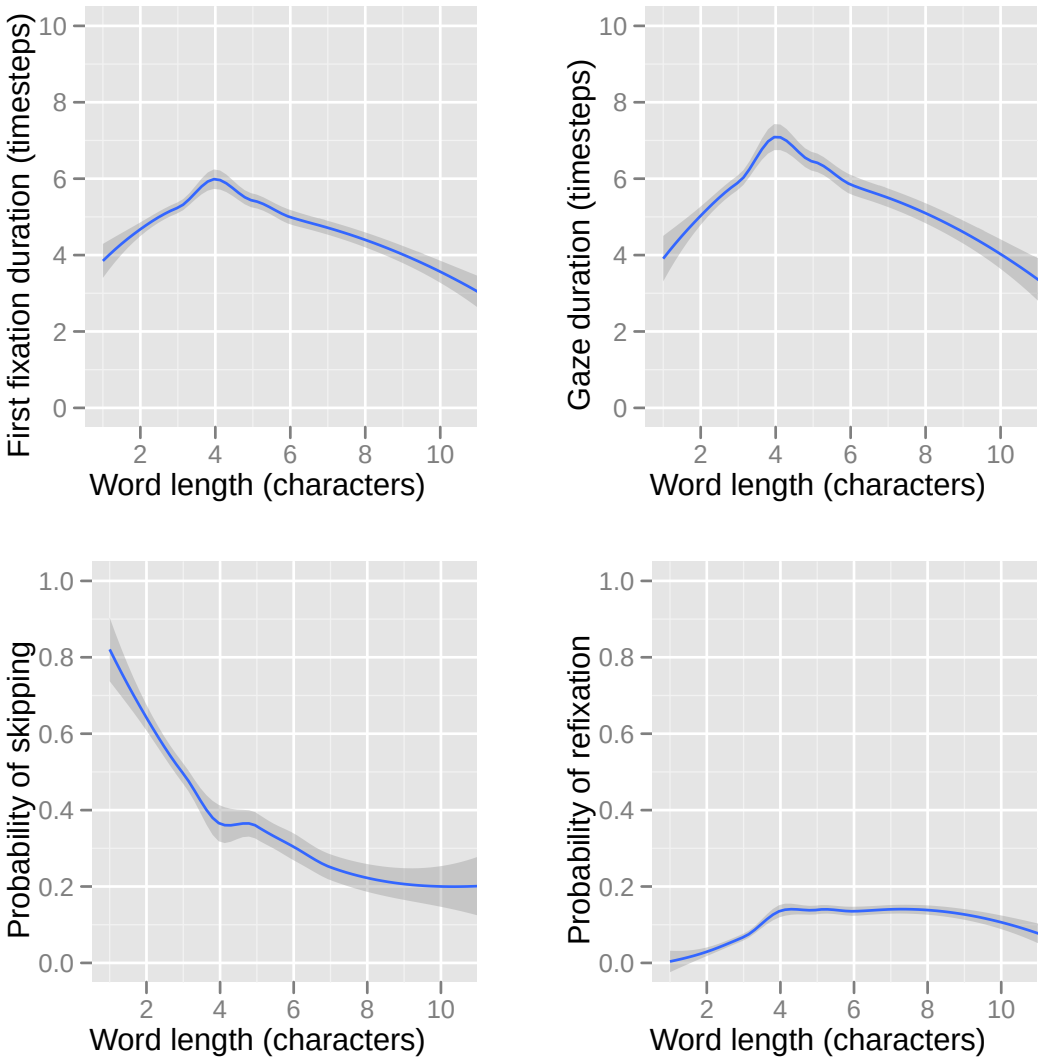

Figure 4.3: The full model's predicted effect of word length on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation, as estimated from simulations by loess, plotted with standard errors calculated across word tokens.

measures, there was some non-monotonicity. The effect of word length on all four aggregate measures was in the same direction as for humans up to length 4. Past that length, it weakened for skip rates and refixation rates, and reversed for the duration measures.

It is perhaps unsurprising that predictability seems to have the most consistent effect, given the large role that predictability plays in the model, and the relatively straightforward predictions made previously. Perhaps more surprising are the non-monotonicities in the predictions for how the fixation duration measures should vary with respect to word frequency and length. In the previous section, we argued that there were two reasons we might expect a reversed word length effect, both related to limitations of the current implementation of the model: (a) our removal of many low frequency words from the grammar and (b) the assumption that the model has veridical knowledge of word length. As described above, there are reasons to believe that these problems may preferentially affect longer words, which could explain why the effect of word length reverses after a length of four. Because length and frequency are strongly correlated, it may be that these problems are also to blame for the non-monotonicity of the effect of frequency. In the following simulation, we investigate this hypothesis by testing whether remedying one of these two problems, namely the removal of many words from the grammar, will mitigate the non-monotonicities in the predictions for frequency and length effects, by using a grammar in which fewer low frequency words have been removed.

4.3 Simulation 2: Model without context

The main goal of Simulation 2 is to explore the hypothesis that removing low frequency words from the model's vocabulary contributed to the non-monotonicities we observed in the effects of word frequency and length on fixation durations. Recall that the initial reason that the vocabulary was severely trimmed was to allow the full model to run simulations at an acceptable speed. Thus, if we simply trimmed the vocabulary less, simulations would be unacceptably slow. For that reason, our strategy here is to simplify the grammar,

which makes the computations simpler and faster to carry out. Thus, we can use a larger vocabulary and still run the model at a similar speed, since the increase in speed brought about by the grammar simplification is of comparable size to the decrease caused by increasing the vocabulary. Specifically, we replace the model's previous grammar (a bigram language model), which made use of linguistic context, with a grammar that includes only word frequency information (a unigram language model), and thus cannot make use of linguistic context. In addition to allowing us to use a larger, more realistic vocabulary, simplifying the grammar allows us to start to answer the question of which of the assumptions of our model are necessary to its predictions: specifically, how do the model's predictions change when it can no longer make use of the linguistic context to help recognize words.

4.3.1 Methods

The model used in this simulation differs in three ways from that used in Simulation 1. First, we replaced the grammar with a unigram language model, which has only frequency information. (Training on the BNC was performed in the same manner as previously.) Second, we increased the size of the vocabulary: instead of including only the most common 500 words in the BNC, we include all words that occur at least 200 times in the BNC (corresponding to a frequency of 2 per million; about 19,000 words). Finally, we also multiplied the visual input rate by 2.5.[7] This was done because the new language model gives much poorer information about the identities of words, and as a result, more visual input is needed on average to reach a similar level of confidence in word identities. Increasing the visual input rate by 2.5 results in the new model taking a similar number of timesteps to read a sentence as the previous model. With the model resulting from these three changes, we ran simulations and calculated aggregate measures exactly as in Simulation 1.

[7] Specifically, we increased Λ from 4 to 10.

4.3.2 Results

Predictability

Because this model does not make use of the linguistic context in identifying words, there can be no actual effect of predictability in the model. Nevertheless, because of the correlations between predictability, frequency, and length, the model can still show apparent effects of predictability. Thus, we report the effect of predictability (determined using the previous language model with context) on our four aggregate measures, estimated as in Simulation 1, in Figure 4.4.

Unsurprisingly, the effects of predictability here are all substantially smaller than in Simulation 1. That said, all are in the same direction, and can serve as a baseline to interpret the effects of predictability produced by the full model.

Frequency

The effects of frequency predicted by the model, estimated as in Simulation 1, are plotted in Figure 4.5. Looking first at the skip rate and refixation rate graphs, we can see that the model's skip rates are substantially lower than humans and refixation rates are substantially higher (whereas the previous model was very close to human data in both cases). Presumably, this is because this model cannot make use of linguistic context, and thus needs to make more fixations. Note that this a very similar situation to that described in Chapter 2, in which the original version of the Mr. Chips model (which could not make use of linguistic context) also had skip rates that were substantially lower than humans, while our extended version (which could make use of linguistic context) had skip rates closer to those of humans. However, note also that in the present results, the shape of the curves seems to match humans even more closely than those produced by the full model (in Figure 4.2). While speculative, such a result could be taken to suggest that, while using linguistic context is crucial to having skip rates of similar magnitude as humans, using a large vocabulary may be crucial in matching the exact shape of the relationship between frequency and

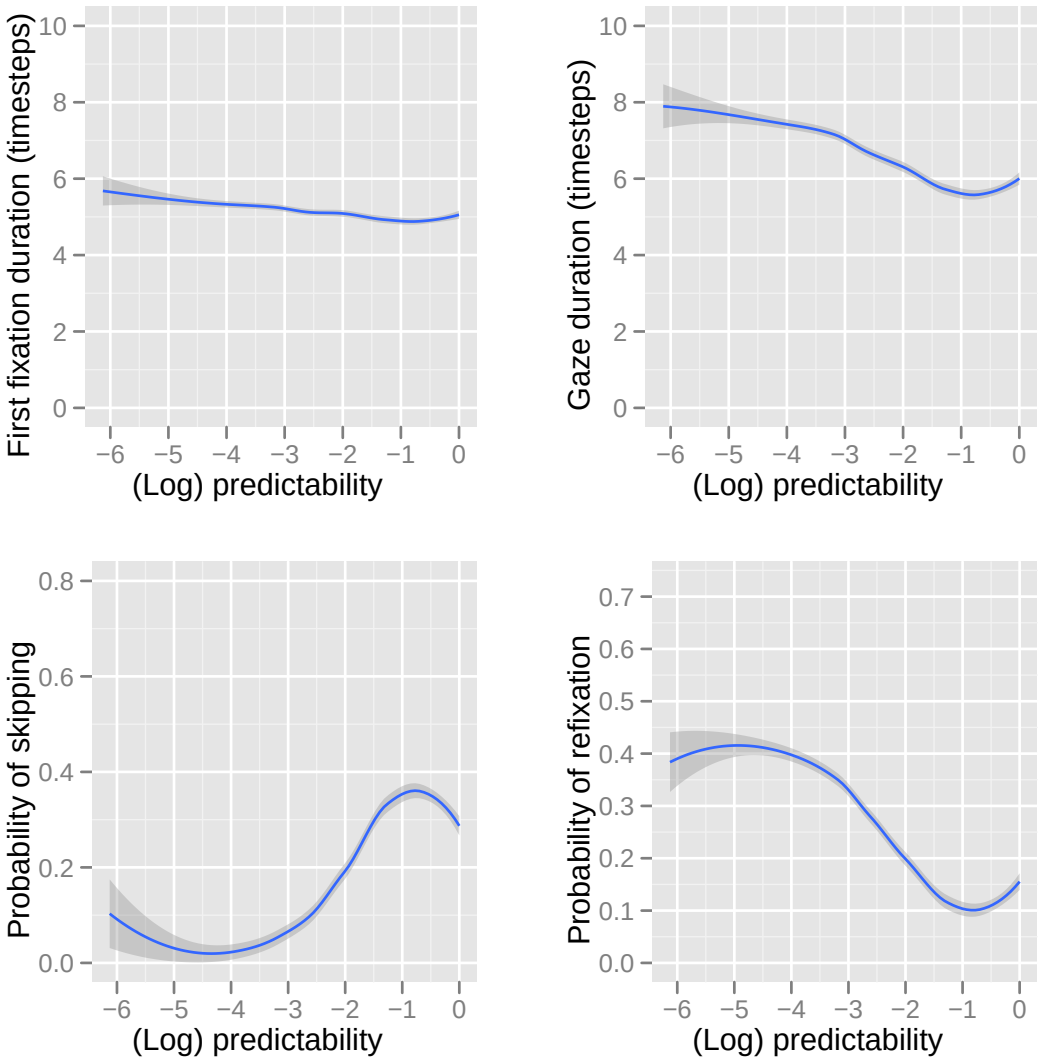

Figure 4.4: The model without context's predicted effect of word predictability on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation, as estimated from simulations by loess, plotted with standard errors calculated across word tokens.

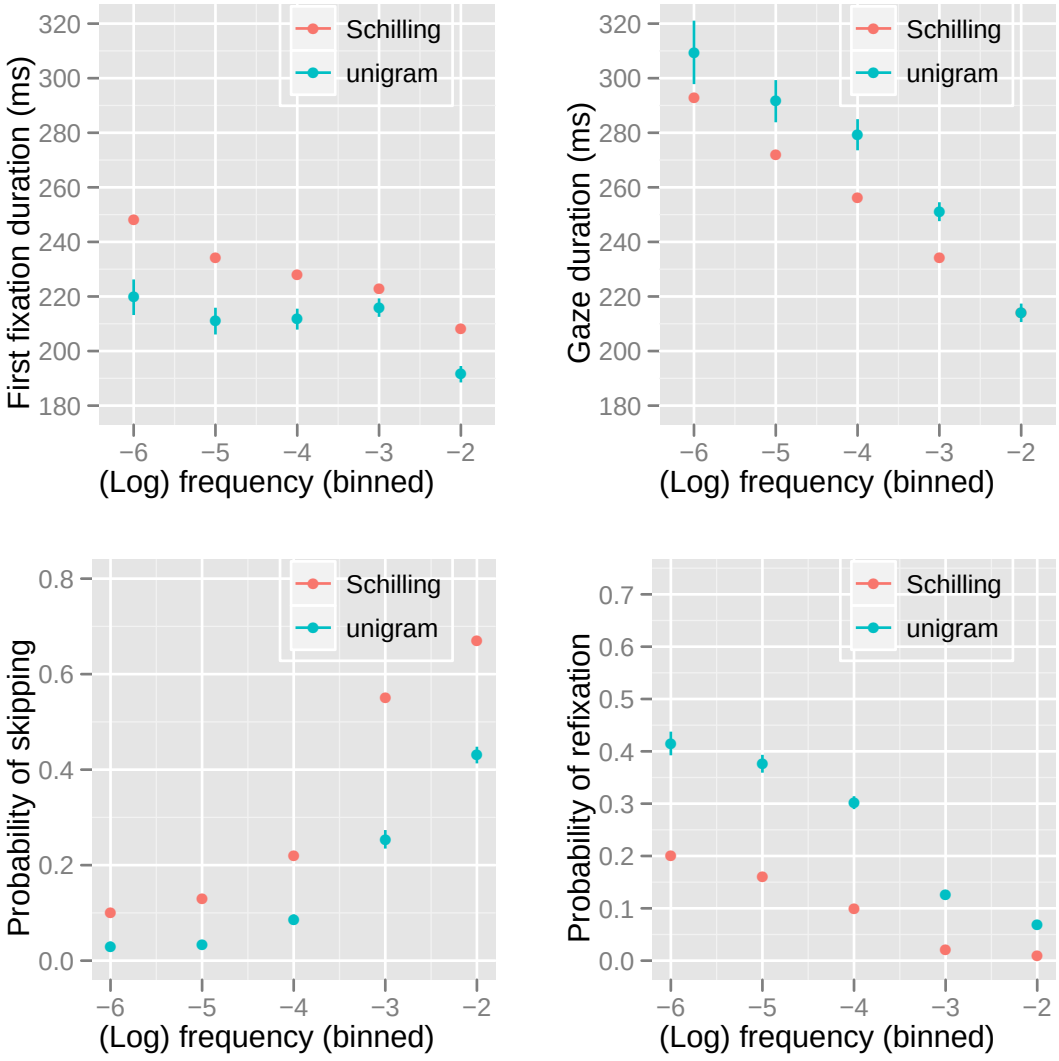

Figure 4.5: The model without context's predicted effects of word frequency on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation, as estimated from simulations, plotted along with standard errors calculated across word tokens. Mean values from the Schilling corpus reported by Pollatsek et al. (2006) are shown for comparison.

skip rates in human data.

Looking now to duration measures, there are two apparent differences between this simulation and the previous one. First, the effect of frequency on first fixation durations now appears to be too small. In addition, the problem of non-monotonicity is now somewhat alleviated: substantially reduced for first fixation durations and completely eliminated for gaze durations. This result supports our argument that trimming the vocabulary may be responsible for some of the non-monotonicity in the previous simulation results.

Length

The effect of word length, estimated as in Simulation 1, is shown in Figure 4.6. As in the previous simulation, the predictions are in the right direction for all four measures up through a word length of about four, at which point – as in the previous simulation – the pattern changes. In this case, the predictions for skipping and refixation rates are in the right direction throughout, while for durations the effect again reverses, although the reversed effect is substantially smaller than before. Thus, the predictions for length effects provide further evidence for the notion that some part of the non-monotonicity was driven by the use of an artificially small vocabulary. The fact that some non-monotonicity still remains suggests that the artificially small vocabulary is not the only source of the problem, however. One further possibility that has already been mentioned is that artificially giving the model veridical knowledge of word length may also take some responsibility for this pattern.

4.3.3 Discussion

Together, the pattern of frequency and word length effects produced by this model with a larger, more realistic vocabulary suggests that at least some of the non-monotonic predictions made by the model in Simulation 1 were caused by its artificially small vocabulary. In addition, these results could be taken to suggest that the ability to use linguistic context may be crucial to producing human-like reading behavior. Specifically, there is a smaller frequency effect for

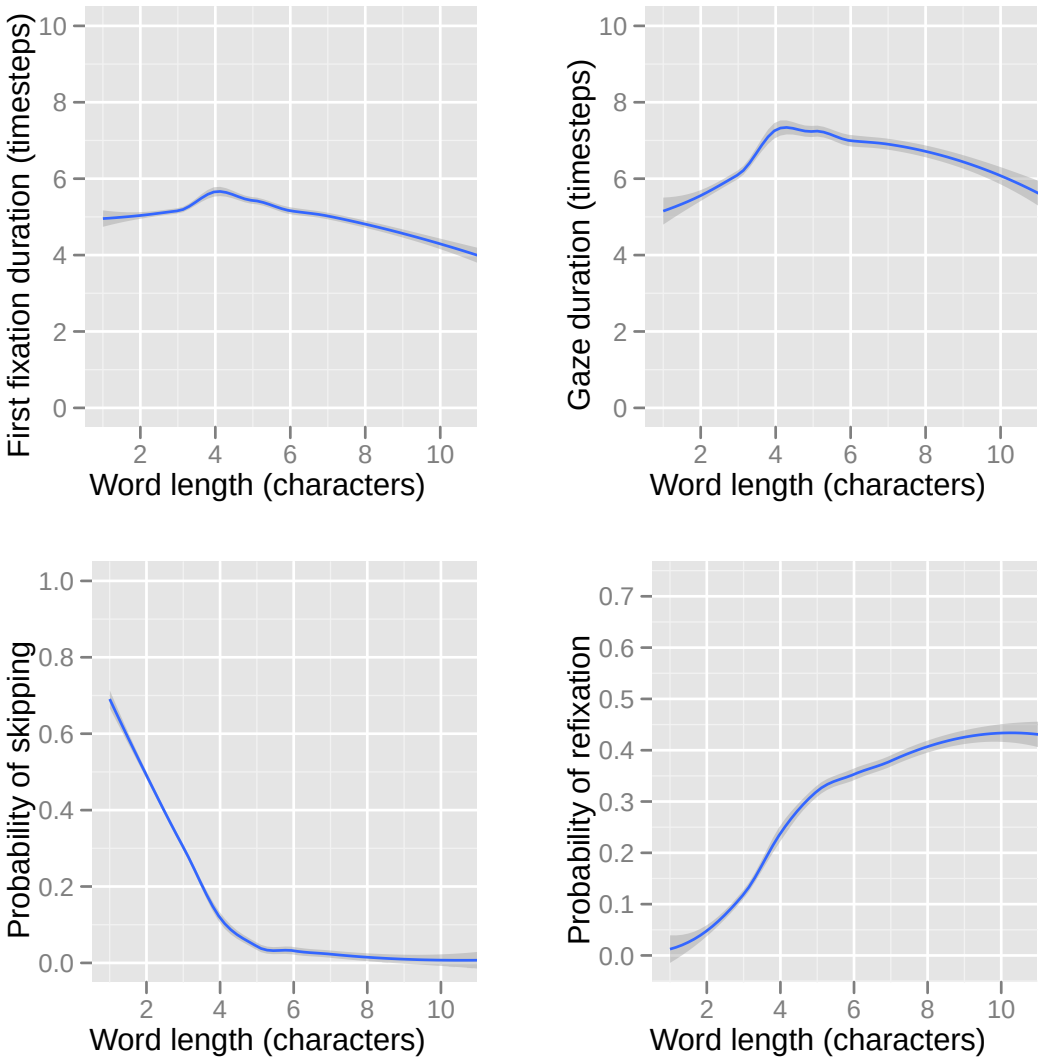

Figure 4.6: The model without context's predicted effect of word length on first fixation durations, gaze durations, the rate of skipping, and the rate of making a refixation, as estimated from simulations by loess, plotted with standard errors calculated across word tokens.

first fixation durations (yet a normal sized one for gaze duration), the refixation rate is far too high, and it virtually never skips any words longer than 4 characters.

4.4 General discussion

In this chapter, we first described why a model in the rational reading framework might be expected to produce the well known effects of word predictability, frequency, and length. We then described the results of two simulations run with models in the rational reading framework testing for effects of these three variables. Simulation 1, using a model similar to that used in the previous chapter, demonstrated that we can derive effects of frequency and predictability from principles of rational inference and a very simple behavior policy (though the predictions for fixation durations for words of intermediate frequency are not completely monotonic). In addition, there is some evidence that we can derive effects of word length, although the effects were only in the right direction for word lengths up to four characters.

We argued that each of the problems with these predictions may be a result of limitations of the model's current implementation. Simulation 2, using a model with a simpler grammar but larger vocabulary, provided some evidence that removing one of the limitations from the current implementation (namely, the artificial restriction on vocabulary size) can help to remedy these shortcomings. In addition, it suggests that the use of context may play an important role in accurately reproducing humans' reading behavior.

The single biggest problem in this pattern of results is that, as already noted, there are inverted effects of word length on fixation durations in both Simulations 1 and 2. When discussing the theoretical predictions for effects of word length, we noted that there are at least two reasons why we might expect such a pattern of results, both stemming from limitations of the current implementation of the model. The first is that it has an artificially small vocabulary and the second is that it has veridical knowledge of word length. Results from Simulation 2 suggest that the artificially small vocabulary is responsible for at

least some of these effects, as the inverted effect becomes smaller when the vocabulary is enlarged. However, the fact that an inverted effect still remains may be an indication that the assumption of veridical knowledge of word length is also contributing. It will be one of the major challenges of future work in this framework to find a way of relaxing this assumption as well, in such a way that the model can still run efficiently.

These results demonstrate that the rational reading framework described in the previous chapter can produce reasonable effects of word predictability, frequency, and length on four aggregate measures of eye movement behavior: first fixation durations, gaze durations, skip rates, and refixation rates. While these predictions are not perfect, it is striking that they are in fair agreement with human readers given that we have made no efforts to tune our model to fit human data. Future work developing the predictions of our model with respect to these basic effects will fall into three lines. First, as already discussed, the results described here suggest that we should work on relaxing the model's simplifying assumption that it has veridical knowledge of word length, in order to improve the model's predictions for the effect of word length. Second, there are a few free parameters in our model, and we have not systematically explored how model behavior changes when they are set to different values. In order to better understand which assumptions of the model are crucial to its predictions, we should systematically explore the effects of manipulating those parameters. Finally, it should be noted that while these four aggregate measures are widely used to summarize reading behavior, they do not tell the whole story. For this reason, we plan to test other aspects of our model's behavior, such as effects on character landing positions within words and on regressive saccades.

Chapter 5

Why readers regress to previous words: A statistical analysis

Klinton Bicknell and Roger Levy

Abstract. While the major models of eye movement control in reading propose very different mechanisms for the generation of saccades to previous words, there has been relatively little empirical data to distinguish these hypotheses. Here we provide a systematic statistical analysis of the factors that elicit these saccades in a corpus of eye movements. We show that the results are contrary to the predictions of a number of accounts, and provide new evidence to discriminate among the models.

5.1 Introduction

The control of the eyes during reading is one of the most complex everyday tasks humans face, as efficient performance requires the rapid integration of complex information from visual, motor, and linguistic sources. In recent decades, our knowledge of the determinants of how readers select saccade targets and decide how long to fixate particular locations has grown immensely (for reviews see Rayner, 1998, 2009). The focus of this work has been on understanding the most common ways in which the eyes move through the text: progressive saccades to a word further in the text and refixations of the current word. As such, our understanding of instances in which the eyes move back to a previous word (between-word regressions) is still one of the most poorly developed facets of theories of eye movement control in reading. While not the most common type of saccade, between-word regressions are still a regular property of the eye movement record, consistently occurring on between 1 in 10 and 1 in 20 saccades (and as high as 30% of all saccades for some readers, Radach & McConkie, 1998). Thus, it is a striking state of affairs that the major models of eye movement control in reading (e.g., Pollatsek et al., 2006; Engbert et al., 2005) propose very different reasons for making between-word regressions, each of which is intimately tied to the model's view of the nature of eye movements in reading. One of the best ways to tease apart the various models of eye movement control, then, is to gain a better understanding of between-word regressions. To date, however, there has been relatively little empirical data analysis to distinguish these various theories of why a reader would move their eyes to a previous word. Thus, the present work seeks to remedy this situation by providing a systematic analysis of the factors that elicit between-word regressions, and in so doing to provide a new source of evidence distinguishing among models of eye movement control in reading.

While there are many proposals as to why a reader would regress to a previous word, it is a common belief among researchers who have diverse opinions on the primary cause of between-word regressions (e.g., Reichle et al., 2009; Engbert et al., 2005) that some proportion of them are made in response

to overshooting a target word due to motor error, and making a between-word regression to return to the originally targeted word. Supporting this view is evidence that between-word regressions increase following word skipping (Vitu & McConkie, 2000). Thus, to better distinguish between competing theories of between-word regressions, we focus our discussion here only on cases in which the target of the between-word regression was not skipped (about half of all between-word regressions in the dataset we examine below). The question for the theories then becomes one of why a reader would regress to a previously fixated word.

5.1.1 Theories of between-word regressions

We consider here five classes of theories of the causes of between-word regressions to unskipped words. For each theory, we highlight the predictions it makes for which variables will influence the rate of between-word regressions. For concreteness, we discuss these predictions for the case of a regression from the nth word in a given sentence (word n) to the previous word $n-1$. We can group the factors to which regressions are predicted to be sensitive into three categories: (a) properties of word n (i.e., its length, frequency, or predictability), (b) those same properties of word $n-1$, and (c) motor properties (i.e., length of the previous saccade or the position of the eyes on word n or previously on word $n-1$).

Corrective

One possibility for most theories is that some between-word regressions (even when word $n-1$ was not skipped) could still be corrective. This could happen if the saccade that landed on word n was intended to be a refixation of word $n-1$. This is quite plausible since the most common pattern of refixations is that first a word's beginning is fixated and then its end (Rayner, Sereno, & Raney, 1996).

This account predicts that properties of word $n-1$ should be relevant, since there should be more regressions when refixations are more likely (lower

frequency or predictability, longer length of word $n-1$). In addition, motor properties should be relevant: regressions should be more likely when the eyes land closer to the beginning of word $n-1$ (since refixations are more likely) and closer to the beginning of word n (since that is where failed refixations would land). Properties of word n are not predicted to matter, except insofar as they correlate with properties of the preceding input.

Oculomotor strategy

Another possibility is that regressive saccades could be initiated as part of an oculomotor strategy (O'Regan & Lévy-Schoen, 1987; Yang & McConkie, 2001). On this account, readers have learned that it is generally beneficial to launch a regression in response to a particular configuration of visuomotor variables. For example, a strategy could be to regress after a particularly long saccade, or after skipping a word. While the precise details of these strategies have not been well worked out for cases in which a word was not skipped, the crucial prediction is that regressions produced by an oculomotor strategy cannot be influenced by linguistic properties like frequency and predictability.[1]

Incomplete lexical processing

A range of models argue that regressions can sometimes be produced by incomplete lexical processing (in reading, generally taken to be synonymous with word recognition). There are two possible accounts of how this could happen. In a serial word processing model, in which readers attend a single word at a time, it could occur when a reader accidentally moves their eyes away from a word too early, and then moves them back to continue processing it more efficiently (Vitu, McConkie, & Zola, 1998).[2] However, it is in attention gradient

[1] Of course, in models such as Yang and McConkie (2001), higher level language processing can sometimes intervene, so that not all regressions in this model would be produced by an oculomotor strategy. For these cases however, the regression must be produced by one of the other accounts described.

[2] Note however that in the most well developed serial model of eye movements in reading, *E-Z* Reader (Reichle et al., 2009), this could not happen because a saccade is only initiated to leave a word after all visual processing is completed.

models, in which readers attend multiple words simultaneously, that this account has been better developed (Engbert et al., 2002, 2005; Reilly & Radach, 2006). There, if the processing of the previous word was too short (relative to the word's length, frequency, and predictability), then its activation can become higher than that of the current word or future words, which can trigger a between-word regression.

Both of these accounts predict that factors that increase the difficulty of word $n-1$ (longer length, lower frequency and predictability) should increase the number of regressions (for a given fixation duration on word $n-1$). In addition, the attention gradient models predict that regressions should be more likely when word n is easier (shorter length, higher frequency and predictability), since its activation level will thus be less of a competitor. The serial model predicts that linguistic properties of word n will not have an effect.

Integration failure

It has been well documented that strong garden path sentences (i.e., sentences with temporarily ambiguous words or syntactic structures, which are initially strongly biased towards the incorrect interpretation) often elicit between-word regressions at the disambiguating region (Frazier & Rayner, 1982). A common explanation for this finding is that integration of the disambiguating word into prior context fails, and readers must regress to previous words for reprocessing. Although most evidence for regressions in this situation comes from experimental manipulations with strong, artificial garden path sentences, it may be that weaker garden paths (which are not consciously perceptible) nevertheless sometimes cause integration failure and elicit regressions through this mechanism. It should be noted, however, that a large number of such garden paths would be required to produce between-word regressions on 5-10% of saccades.

Since the difficult disambiguation region in a garden path is by definition unpredictable, this account predicts more regressions when word n is less predictable (the opposite of the prediction made by the attention gradient incomplete lexical access account). In addition, it has often been found that garden

path regressions occur on following words, thus this account also predicts more regressions on word n when word $n-1$ is less predictable. It would not obviously predict the length of either word or the eyes' landing position on either word to be relevant.

Confidence falling

The final theory of between-word regressions we discuss here was suggested recently by Bicknell and Levy (2010b). In this account, readers maintain uncertainty about the identities of previous words and update that uncertainty as input from new words further downstream becomes available (Levy, 2008), a proposition that has some recent empirical support (Levy, Bicknell, et al., 2009). The model proposes that when a new word fits relatively poorly with what the reader believed the prior context to be, and relatively better with an alternative visually similar possibility, the reader's confidence in the identity of the prior context will be reduced. In this situation, it becomes useful to get more visual information about the prior context, and thus make a between-word regression.

The predictions of this account combine predictions of the incomplete lexical processing and integration failure accounts. Confidence is more likely to fall about words whose confidence was lower to begin with, which predicts that factors that slow processing of word $n-1$ (longer length, lower predictability) will increase regressions, as in the incomplete lexical processing account. Like the integration failure account, however, this account predicts that an unpredictable word n (since it fits poorly with the prior context) will be more likely to cause confidence to fall, and the word's length should be irrelevant. The prediction for predictability is actually more subtle, however, since not every word that is unpredictable given a particular context will be more predictable given some other context. We will return to this point later.

5.1.2 Previous empirical evidence

There is relatively little empirical evidence regarding which factors between-word regressions are sensitive to in the case that the regression tar-

get word was not skipped. Most of the existing work (which has not generally controlled for skipping) has looked for effects of linguistic properties of the targets of regressions. For example, readers were found to regress more to words of low predictability with no significant additional effect of frequency (Rayner, Ashby, Pollatsek, & Reichle, 2004; Kliegl, Grabner, Rolfs, & Engbert, 2004). However, these studies failed to control for skipping, and thus the results could be confounded, if, for example, regressions are triggered more towards words that were unintentionally skipped. The one study that has investigated the effects of properties of word $n-1$ on between-word regressions specifically in the case in which word $n-1$ wasn't skipped found that there were more between-word regressions when word $n-1$ was lower frequency and longer (Vitu & McConkie, 2000). However, Vitu and McConkie did not have predictability in their model, and thus could not distinguish between effects of predictability and frequency. More crucially, however, their analysis could not determine whether there were independent effects of frequency and length (since longer words are also generally less frequent). Thus, while there is good evidence that some properties of the regression target are implicated for the case in which the regression target was not skipped, it is not clear whether it is the length, frequency, or predictability of the word that attracts regressions.

Evidence about whether properties of word n make regressions more likely is still more scarce. The one reported result is that Kliegl et al. (2004) found that words that were of lower frequency and lower predictability were more likely to have a regressive saccade begin on them. However, to the best of our understanding Kliegl et al. did not control for whether a word was fixated at all, and thus these results are confounded with word skipping, since a word that was skipped could by definition not have a regression begin on it.

The evidence to date for effects of motor properties on between-word regressions also comes from Vitu and McConkie (2000). They found that, contrary to the overall tendency for regressions to be more likely following longer saccades (Buswell, 1920; Vitu et al., 1998), in the case that word $n-1$ was not skipped, regression rates decrease with following longer saccades. In addition, Vitu and McConkie reported a non-significant trend for regressions to decrease

as word $n-1$ was fixated further from its beginning.

Thus, while Vitu and McConkie's (2000) results demonstrate that cases in which word $n-1$ was not skipped pattern very differently from cases in which it was, it is presently far from clear what the determinants of between-word regressions are in this condition. There is evidence that properties of word $n-1$ are relevant, but it is unclear which ones, and we know almost nothing about whether properties of word n are relevant. As noted above, many of the theories of regressions make predictions for factors such as the position of the eyes in word $n-1$ or word n, yet a reliable effect of such variables has not been found. As a result, all five classes of theories mentioned above are still quite tenable explanations for making a regression to an unskipped word. Finally, as pointed out by Vitu and McConkie, it is important to realize that many of these factors are highly correlated with one another (for example, the landing position within a word and its length and frequency), and thus strong evidence that a variable is relevant for regressions can only be made using a model controlling for effects of correlated factors. Thus, the goal of the analysis reported in this paper was to simultaneously and systematically test for effects of a range of variables – including properties of words n and $n-1$ as well as motor variables – on the rate of between-word regressions in cases in which the regression target had not been skipped.

5.2 Method

5.2.1 Corpus and dataset

In this analysis we modeled the rate of between-word regressions in a large corpus of eye movements in reading, the Dundee corpus (Kennedy & Pynte, 2005). This corpus is comprised of the eye movement record of 10 individuals each reading 50,000 words of British newspaper editorials. In order to have a more controlled dataset, we focus only on predicting regressions from word n to the previous word $n-1$ (which account for about 70% of between-word regressions in our dataset), so that no other words intervene, and (for

reasons mentioned above) only in the case that word $n-1$ was not initially skipped. Specifically, we predict whether each saccade that originated on some word n was a regression or not, in the case that (a) the previous saccade originated on word $n-1$, (b) neither words n nor $n-1$ were previously fixated, (c) no word beyond n was previously fixated (i.e., first pass reading), (d) neither the previous fixation (on word $n-1$) nor the next fixation following the saccade in question were the first or last on a line nor detected as a blink, and (e) the saccade was not a regression going back further than word $n-1$. Each saccade meeting these criteria was thus categorized as a regression if it went to word $n-1$, or a non-regression if was a refixation of word n or a progressive between-word saccade. Finally, we excluded cases in which words n or $n-1$ were not in the British National Corpus (see below), had punctuation (including all non-alphabetic characters), or were the first or last words in a line, as well as any case in which the fixation on word n or $n-1$ had been on the space preceding the word. This resulted in a dataset of 33569 saccades, of which 1362 or 4% were regressions.[3]

5.2.2 Analysis

We fit a generalized linear mixed-effects regression with a logit link function (Pinheiro & Bates, 2000; Agresti, 2002) to the data using the `lme4` package (Bates & Maechler, 2010). The fixed effects in the model included the factors discussed above: properties of words n and $n-1$ (their log-transformed frequency, predictability, and length) as well as motor properties (the log-transformed length of the previous saccade and the landing positions on both words).[4] In addition, the model included fixed effects for the length of the fixations on both words and random intercepts, but not random slopes, for each par-

[3] This is lower than the overall rate of 5-10% mentioned previously because of the exclusion of between-word regressions going back further than word $n-1$.

[4] The length of word $n-1$, the landing positions on both words, and the saccade length form a linearly dependent set such that the fourth is completely determined given the values of the other three, and thus including these four variables in a single model directly would be impossible. In our case, this is not a problem because three of the four are log-transformed, removing the linear dependence. The fact that multi-collinearity still exists between them, however, means that the estimates of their effects may be conservative.

ticipant. (Models including random slopes for the nine predictors of interest failed to converge.) Frequency and predictability were estimated by unigram and trigram language models trained on the British National Corpus, smoothed with modified Kneser-Ney smoothing (Chen & Goodman, 1998). Because coefficient estimates in models without random slopes for participants can be anti-conservative for datasets in which there is real between-participant variability in effect sizes, we performed statistical tests by bootstrapping instead of using the standard normal-theory statistics (Efron & Tibshirani, 1993).[5] Specifically, we obtained p-values and 95% confidence intervals for each coefficient from 2500 replicates of hierarchical bootstrapping, clustered by participant (Davison & Hinkley, 1997).[6]

5.3 Results

The marginal effects of the properties of words n and $n-1$ are plotted in Figure 5.1. The results of the regression reveal significant effects of all three properties of word $n-1$: regressions were more frequent when it was longer, more frequent, and less predictable (ps $< .0008$). Regressions also increased when word n was less predictable ($p < .0008$), but were not sensitive to its length or frequency (ps $> .3$). Finally, regressions were less likely as the length of the previous saccade increased ($p < .02$) and as the landing position on word $n-1$ was further from the beginning ($p < .01$). The relative contributions of each of these factors to the likelihood of a regression in the full model is visualized in Figure 5.2.

[5] Bootstrapping in this case also avoids the potential problems that normal-theory statistics are not completely valid when using the Laplace approximation to the model likelihood surface and that the Wald test becomes conservative when the data are very near to 0 or 1, as is the case for our dataset.

[6] We denote by $p < .0008$ cases in which the estimate of a coefficient in all bootstrap replicates is on the same side of zero, since if a single replicate had been on the other side of zero, the probability would be twice 1/2500, or .0008.

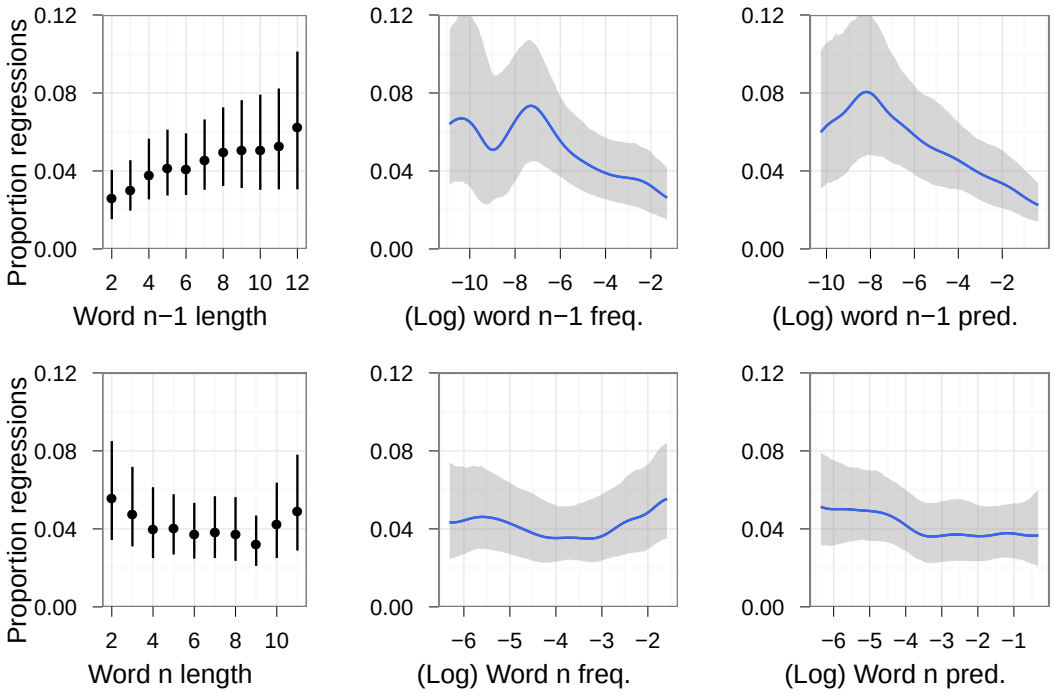

Figure 5.1: Marginal effects of length, frequency, and predictability of words n and $n-1$ on proportion of regressions to word $n-1$, shown for the middle 95% of the range of each variable. Proportion of regressions was estimated using Gaussian kernel regression with standard deviation equal to 1/15th of this range. The 95% confidence intervals are hierarchically bootstrapped from 1000 dataset replicates (Efron & Tibshirani, 1993).

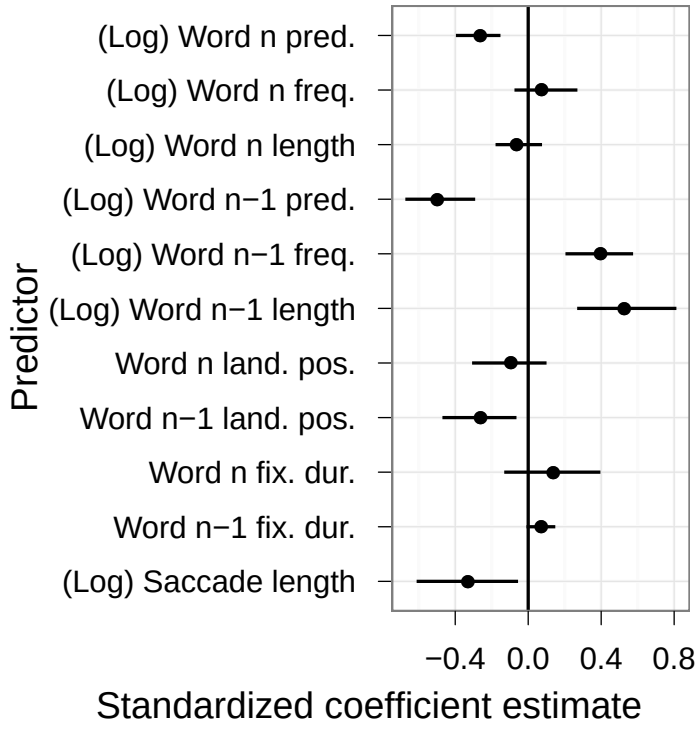

Figure 5.2: Estimates and 95% confidence intervals of the predictor coefficients, standardized to be on the same scale to visualize the relative contributions of each factor to the full model. (Standardization was performed by multiplying the actual coefficient by the standard deviation of the predictor.)

5.3.1 Additional analysis

Prior to discussing the implications of these results for the theories under discussion, we note that having frequency and predictability effects in opposite directions for word $n-1$ (i.e., increased regressions for less predictable but more frequent words) is not an obvious prediction of any theory. The graphs in Figure 5.1, however, show that the marginal effects of frequency and predictability for both words are in the same direction (increased regressions for less predictable and less frequent words) when not both in the model together. This pattern appears in the marginal effects because of the high correlations between frequency and predictability ($r = .89$ and $r = .73$ for words $n-1$ and n). Despite this, the regression model reports highly significant effects of the predictability and frequency of word $n-1$ in opposite directions.

This pattern of opposite effects of frequency and predictability may be well understood in terms of the confidence falling account. As noted above, this account predicts that not every unpredictable word should cause more regressions, but only unpredictable words which are more predictable for alternate possible contexts, since they cause confidence about the true context to fall. Thus, one rough measure of the amount by which a word causes a reader's confidence to change about the preceding context (denoted Δ_c) is the change in log probability of the true context after seeing the word, relative to the context's overall likelihood:

$$\Delta_c = \log p(c|w) - \log p(c)$$

Thus, we analyzed a new model in which we replaced predictability for both words with this measure of the change in confidence about the prior context given each word. It turns out, in fact, that this measure is equivalent to the difference between the word's log-transformed frequency and predictability

$$\begin{aligned}\Delta_c &= \log p(c|w) - \log p(c) = \log \frac{p(w|c)p(c)}{p(w)} - \log p(c) \\ &= \log p(w|c) - \log p(w)\end{aligned}$$

and because of this relationship, the new model is simply a reparameterization of the former model (i.e., the fit and predictions will be identical), in which the

coefficient for Δ_c is identical to the previous coefficient for predictability and the new coefficient for the additional effect of frequency is the sum of the previous coefficients for predictability and frequency, as can be seen in the following equation (where f and p denote frequency and predictability):

$$\beta_1 p + \beta_2 f = \beta_1(p - f) + (\beta_1 + \beta_2)f$$

(All other coefficient values will remain the same.)

In this new model, the effect of Δ_c is such that there are significantly more regressions the further confidence falls on either word (ps $< .0008$).[7] There are no additional effects of the frequency of either word (ps $> .1$), providing support to the notion that confidence falling may be a useful way to understand the opposite effects of predictability and frequency.

5.4 Discussion

We discuss the implications of this pattern of results for each of the five classes of theories separately.

Corrective

It seems unlikely that corrective saccades are driving any effects for this dataset because of the directionality of the effect of landing position on word $n-1$. As mentioned above, landing positions closer to the beginning of word $n-1$ should result in more attempted refixations of the word's end, and thus more unintentional fixations on the early part of word n. Furthermore, of course, the corrective account could not predict the effect of the predictability of word n.

Oculomotor strategy

It is similarly unclear how an oculomotor strategy could account for these findings, since the linguistic properties of both words have strong effects.

[7] This is of course necessarily the case, since Δ_c has the same coefficient as predictability had previously.

Also, we note that the particular strategy of making regressions after especially long saccades directly conflicts with our data.

Incomplete lexical processing

The incomplete lexical processing account correctly predicts the effects in our data of predictability and length of word $n-1$. As noted above, however, these accounts predict either that regressions should not be sensitive to properties of word n (serial models) or that they should increase when word n is easier (gradient models). Thus, the finding that regressions increase when word n is less predictable poses a problem for these accounts.

Integration failure

The integration failure account is supported by the fact that more regressions are made as either word becomes less predictable. The fact that the length of word $n-1$ also had a significant effect is problematic. One possible explanation for this is that some portion of the effect we obtained may be an artifact of the way we constructed our dataset. Specifically, by considering only between-word regressions made to word $n-1$, we may have excluded more regressions which targeted word $n-1$ but which landed on a word prior to $n-1$ in the case that word $n-1$ was shorter.[8]

Confidence falling

The confidence falling account correctly predicts the effects of the predictability of words n and $n-1$ as well as the effect of the length of word $n-1$. As revealed by the additional analysis including Δ_c, the pattern of predictability and frequency effects that was found can be interpreted as some evidence

[8] To investigate this possibility, we performed a similar analysis predicting all intra-line between-word regressions, and not only those to word $n-1$. The results showed the effect of the length of word $n-1$ to be marginal ($p = .07$), supporting the notion that our censored dataset may be responsible for a large part of the effect, but still hinting that the relationship may exist apart from censoring.

that the amount by which a word makes confidence fall is a key determinant in whether a reader will make a regressive saccade.

5.5 Conclusion

We distinguished five classes of models of regressions to previously fixated words, and tested these accounts by performing a systematic statistical analysis of such regressions in a large eye movement corpus. The results of our analysis provide some of the clearest evidence to date about the variables contributing to between-word regressions. The analysis reveals strong effects of linguistic properties, and thus are hard to accommodate in purely corrective or oculomotor strategy accounts. In addition, the fact that there are more regressions to the previous word when the current word is less predictable is counter to the predictions of incomplete lexical processing models. Both integration failure and confidence falling accounts are consistent with the present data, but the facts that (1) the length of the previous word appears to matter and that (2) the opposing effects of frequency and predictability can be understood as falling confidence, suggest that the confidence falling account may find more support in the present data.

More generally, our results demonstrate that obtaining more detailed knowledge of the factors contributing to between-word regressions can distinguish between models of eye movements in reading which otherwise make very similar predictions for progressive saccades. Specifically, SWIFT (Engbert et al., 2005) makes regressions via the incomplete lexical processing account, which appears to make the wrong predictions for our results, while *E-Z* Reader 10 (Reichle et al., 2009) makes use of integration failures, which is consistent with our data. Finally, we note that the empirical success of the confidence falling account, which follows from a very different class of reading model than the others considered, suggests that gaining a better understanding of regressions may have important consequences for our understanding of eye movement control in reading in general.

5.6 Acknowledgements

Chapter 5, in full, is an exact copy of the material as it appears in Bicknell and Levy (2011) [Why readers regress to previous words: A statistical analysis. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.] 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 24th Annual CUNY Conference on Human Sentence Processing.

Chapter 6

Conclusion

This dissertation presented a rational framework for understanding eye movements in reading and provided arguments and empirical evidence in support of this framework. Chapter 2 presented the first rational model of eye movements in reading, Mr. Chips, and described an extension of it, which added to the model knowledge that humans arguably also use. It highlighted the ability of the Mr. Chips model to replicate human word skipping rates, initial fixation locations on words, and refixation rates, and also provided evidence that the extension made the model's predictions even closer to human data. Chapter 3 presented a new rational model of eye movements in reading and used it to demonstrate that sometimes making regressions to a previous word is a natural consequence of efficient reading. Chapter 4 described the reasons why this new model should produce the well-known effects of three linguistic variables – word predictability, frequency, and length – and also demonstrated that simulations run with the model do produce these effects. Finally, Chapter 5 provided empirical support for the new model's account of regressive saccades, in fact demonstrating that it is the only single account of regressive saccades that can account simultaneously for all the linguistic effects observed in that analysis.

Together, these results suggest that we can understand a number of properties of human eye movements during reading as arising naturally as part of an efficient solution to the task. This framework not only provides new insight into the reasons for well-known effects – such as the linguistic variables of word

predictability and frequency – but also makes new predictions for poorly understood phenomena such as between-word regressions. In addition to providing insights into the reasons for these linguistic effects, the rational framework makes specific the link between reading behavior and language processing, and should help language researchers to design more specific predictions regarding the effects that linguistic manipulations will have on the eye movement record.

This line of research is currently moving in a number of directions. Some of the current work relates to evaluation, i.e., examining more aspects of the model's behavior, such as its predictions for word landing positions. Other current work includes removing the model's veridical knowledge of word boundary information, which – as discussed in Chapter 4 – may be crucial to making the correct predictions for effects of word length. In addition, we plan to extend the model in a number of ways in order to make new predictions. For example, replacing the model's current word-based language knowledge with one that includes syntactic information should allow the framework to make predictions for effects of syntactic manipulations on eye movement behavior. We are also exploring the effect of using other classes of behavior policies, such as some analogous to that used by the Mr. Chips model, to determine the role of the particular behavior policy we used in making the model's predictions.

Finally, we note that because the rational framework is inherently goal-directed, it is well suited to modeling differences in reading behavior across different types of reading tasks, from the typical reading for meaning, to skimming for a particular piece of information, to single word identification, to careful reading, proofreading, and more. These types of changes in task would amount to a change in goal for the model. Changing the goal would in turn alter the type of optimal behavior policy the model uses. The ultimate goal with this framework is not only to provide a principled explanation and model of why reading behavior looks the way it does, but to provide a unified theory of reading behavior spanning all of these domains.

References

Agresti, A. (2002). *Categorical data analysis* (Second ed.). New York, NY: John Wiley & Sons.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., & Mohri, M. (2007). OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)* (Vol. 4783, p. 11-23). Springer.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bates, D. M., & Maechler, M. (2010). *lme4: Linear mixed-effects models using S4 classes.* R package version 0.999375-37.

Bicknell, K., & Levy, R. (2010a). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1142–1147). Austin, TX: Cognitive Science Society.

Bicknell, K., & Levy, R. (2010b). A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL* (pp. 1168–1178). Uppsala, Sweden: Association for Computational Linguistics.

Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*, 1–12.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Buswell, G. T. (1920). *An experimental study of the eye-voice span in reading.* Supplementary Educational Monographs, No. 17. Chicago: University of Chicago.

Chen, S. F., & Goodman, J. (1998). *An empirical study of smoothing techniques for language modeling* (Tech. Rep. No. TR-10-98). Cambridge, MA: Computer Science Group, Harvard University.

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, *30*, 234–250.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. New York: Cambridge University Press.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*, 641–655.

Engbert, R., & Krügel, A. (2010). Readers use Bayesian estimation for eye movement control. *Psychological Science*, *21*, 366–371.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42*, 621–636.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777–813.

Engel, G. R., Dougherty, W. G., & Brian Jones, G. (1973). Correlation and letter recognition. *Canadian Journal of Psychology*, *27*, 317–326.

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

Genzel, D., & Charniak, E. (2002, July). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 199–206). Philadelphia: Association for Computational Linguistics.

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In M. Collins & M. Steedman (Eds.), *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 65–72). Sapporo, Japan: Association for Computational Linguistics.

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, *22*, 487–490.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). New Brunswick, NJ: Association for Computational Linguistics.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*. doi:10.1016/j.cogpsych.2010.02.002.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In D. Lin & D. Wu (Eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 317–324). Barcelona, Spain: Association for Computational Linguistics.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*, 153–168.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

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*, 262–284.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

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*, 2219–2234.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: an Ideal-Observer model of reading. *Psychological Review*, *104*, 524–553.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 234–243). Honolulu, Hawaii: Association for Computational Linguistics.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 21086–21090. (Correction in: *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 5260)

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 849–856). Cambridge, MA: MIT Press.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 937–944).

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*, 648–652.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, *23*, 269–311.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Science*, *434*, 387–391.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*, 1–14.

Narayanan, S., & Jurafsky, D. (2001). A Bayesian model predicts human parse preference and reading time in sentence processing. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14* (pp. 59–65). Cambridge, MA: MIT Press.

Ng, A. Y., & Jordan, M. (2000). PEGASUS: A policy search method for large MDPs and POMDPs. In *Uncertainty in Artificial Intelligence, Proceedings of the Sixteenth Conference* (pp. 406–415).

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327–357.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, *116*, 207–219.

O'Regan, J. K., & Lévy-Schoen, A. (1987). Eye-movement strategy and tactics in word recognition and reading. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 363–383). Hillsdale, NJ: Lawrence Erlbaum Associates.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, USA*, *108*, 3526–3529.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed effects models in S and S-Plus*. New York: Springer Verlag.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*, 1–56.

Radach, R., & McConkie, G. W. (1998). Determinants of fixation positions in words during reading. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 77–100). Amsterdam: Elsevier.

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

Rayner, K. (2009). The 35th Sir Frederick Bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*, 1457–1506.

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, *30*, 720–732.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research*, *16*, 829–837.

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*, 1188–1200.

Reichle, E. D., & Laurent, P. A. (2006). Using reinforcement learning to understand the emergence of "intelligent" eye-movement behavior during reading. *Psychological Review*, *113*, 390–408.

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

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*, 4–22.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*, 445–526.

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*, 1–21.

Reilly, R. G., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, *7*, 34–55.

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

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Stankiewicz, B. J., Legge, G. E., Mansfield, J. S., & Schlicht, E. J. (2006). Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 688–704.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends in Cognitive Sciences*, *12*, 291–297.

Vitu, F., & McConkie, G. W. (2000). Regressive saccades and word perception in adult reading. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 301–326). Amsterdam: Elsevier.

Vitu, F., McConkie, G. W., & Zola, D. (1998). About regressive saccades in reading and their relation to word identification. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 101–124). Amsterdam: Elsevier.

Vul, E., Frank, M., Alvarez, G., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference

in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1955–1963).

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*, 598–604.

Yang, S.-N., & McConkie, G. W. (2001). Eye movements during reading: a theory of saccade initiation times. *Vision Research*, *41*, 3567–3585.