

# A Rational Model of Eye Movement Control in Reading

Klinton Bicknell and Roger Levy

Department of Linguistics

University of California, San Diego

9500 Gilman Dr, La Jolla, CA 92093-0108

{kbicknell, rlevy}@ling.ucsd.edu

## 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.

## 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, Pollatsek, & Rayner, 2006; Engbert, Nuthmann, Richter, & Kliegl, 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 2, we briefly describe the leading models of eye movements in reading, and in Section 3, we describe how these models account for between-word regressions and the intuition behind our model's account of them. Section 4 describes the model and its implementation and Sections 5–6 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.

## 2 Models of eye movements in reading

The two most successful models of eye movements in reading are *E-Z Reader* (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle et al., 2006) and *SWIFT* (Engbert, Longtin, & Kliegl, 2002; Engbert et al., 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, Klitz, & Tjan, 1997; Legge, Hooven, Klitz, Mansfield, & Tjan, 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, 2010). 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 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 re-

gressive 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, Warren, & McConnell, 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, Slattery, & Rayner, 2009). Thus, it is possible that later parts of a sentence can cause a reader’s confidence in the identity of the previ-

ous 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.

#### 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 (4.1), visual input (4.2), formalization of the Bayesian inference problem (4.3), control policy (4.4), and finally, implementation of the model using weighted finite state automata (4.5).

#### 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  $i$ th 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  $\ell_i$ . On the next time step, visual input is obtained around  $\ell_i$  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  $\ell_i$  is normally distributed around the intended target  $t_i$  with standard deviation given by a linear function of the intended distance<sup>1</sup>

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

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

#### 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 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

<sup>1</sup>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.

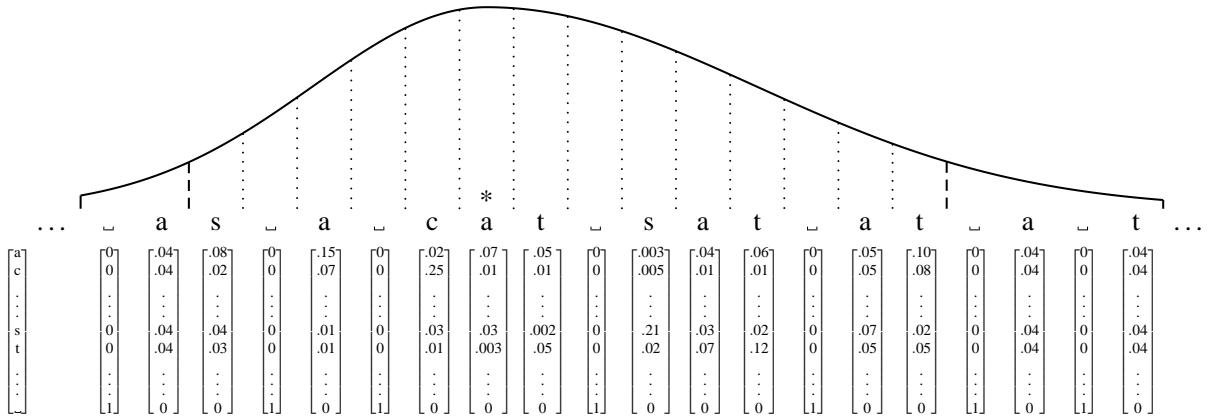


Figure 1: Peripheral and foveal visual input in the model. The asymmetric Gaussian curve indicates declining perceptual acuity centered around the fixation point (marked by \*). The vector underneath each letter position denotes the likelihood  $p(\mathcal{I}(j)|w_j)$  for each possible letter  $w_j$ , taken from a single input sample with  $\Lambda = 1/\sqrt{3}$  (see vector at the left edge of the figure for key, and Section 4.2). In peripheral vision, the letter/whitespace distinction 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.

Figure 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  $\varepsilon$  denotes a character’s eccentricity in characters from the center of fixation, then the proportion of the total processing rate at that eccentricity  $\lambda(\varepsilon)$  is given by integrating the asymmetric Gaussian over a character width centered on that position,

$$\lambda(\varepsilon) = \int_{\varepsilon-0.5}^{\varepsilon+0.5} \frac{1}{Z} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx, \sigma = \begin{cases} \sigma_L, & x < 0 \\ \sigma_R, & x \geq 0 \end{cases}$$

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.

#### 4.2.1 Peripheral visual input

In our model, any eccentricity with a processing rate proportion  $\lambda(\varepsilon)$  at least 0.5% of the rate proportion for the centrally fixated character ( $\varepsilon \in [-7, 12]$ ), 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 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*.

#### 4.2.2 Foveal visual input

In addition, for those eccentricities with a processing rate proportion  $\lambda(\varepsilon)$  that is at least 1% of the total processing rate ( $\varepsilon \in [-5, 8]$ ) the model receives foveal visual input, defined only for letters<sup>2</sup> 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, & 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

<sup>2</sup>For white space, the model is already certain of the identity because of peripheral input.

mean equal to the letter’s true identity and a diagonal covariance matrix  $\Sigma(\varepsilon) = \lambda(\varepsilon)^{-1/2}I$ . 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  $\lambda(\varepsilon)$ . We scale the overall processing rate by multiplying each rate by  $\Lambda$ . 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 1.

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

$$p(w|\mathcal{I}_1^i) = \frac{p(w)p(\mathcal{I}_1^i|w)}{\sum_{w'}(w')p(\mathcal{I}_1^i|w')}. \quad (2)$$

If we let  $\mathcal{I}(j)$  denote the input received about character position  $j$  and let  $w_j$  denote the  $j$ th 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

$$p(\mathcal{I}_1^i|w) = \prod_{j=1}^n \prod_{t=1}^i p(\mathcal{I}_t(j)|w_j). \quad (3)$$

If the eccentricity of the  $j$ th character on the  $t$ th timestep  $\varepsilon_t^j$  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  $\varepsilon_t^j \in [-5, 8]$ , it is the probability of sampling  $\mathcal{I}_t(j)$  from the multivariate Gaussian  $\mathcal{N}(w_j, \Lambda\Sigma(\varepsilon_t^j))$ .

### 4.4 Control policy

The model uses a simple policy to decide between actions based on the marginal probability  $m$  of the

- (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 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  $\beta = .5$ .

most likely character  $c$  in position  $j$ ,

$$\begin{aligned} m(j) &= \max_c p(w_n = c|\mathcal{I}_1^i) \\ &= \max_c \sum_{w':w'_n=c} p(w'|\mathcal{I}_1^i). \end{aligned} \quad (4)$$

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 2. If the value of this statistic for the current position of the eyes  $m(\ell_i)$  is less than a parameter  $\alpha$ , the model chooses to continue fixating the current position (2a). Otherwise, if the value of  $m(j)$  is less than  $\beta$  for some leftward position  $j < \ell_i$ , the model initiates a saccade to the closest such position (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  $m(j) < \alpha$  (2c).<sup>3</sup> Finally, if no such positions exist to the right, the model stops reading the sentence (2d). Intuitively, then, the model reads by making a rightward sweep to bring its confidence in each character up to  $\alpha$ , but pauses to move left if confidence in a previous character falls below  $\beta$ .

### 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

<sup>3</sup>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.

the possible characters for  $w_1$  along with their respective likelihoods given the visual input (as in the inner term of Equation 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.

## 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  $\alpha$  and  $\beta$  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  $\alpha$  but a non-zero  $\beta$ . 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  $\alpha$  and positive  $\beta$ . Our prediction is that the former set will be strictly dominated in terms of both speed and accuracy by the latter.

### 5.1 Methods

#### 5.1.1 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  $\beta \in \{.4, .7\}$ .<sup>4</sup>

#### 5.1.2 Language model

Our reader’s language model was an unsmoothed bigram model created using a vocabulary set con-

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

sisting of the 500 most frequent words in the British National Corpus (BNC) as well as all the words in our test corpus. From this 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).

#### 5.1.3 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.

#### 5.1.4 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.

### 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. As shown in the graph, for each non-regressive policy (the circles), there is a regressive policy that outperforms 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

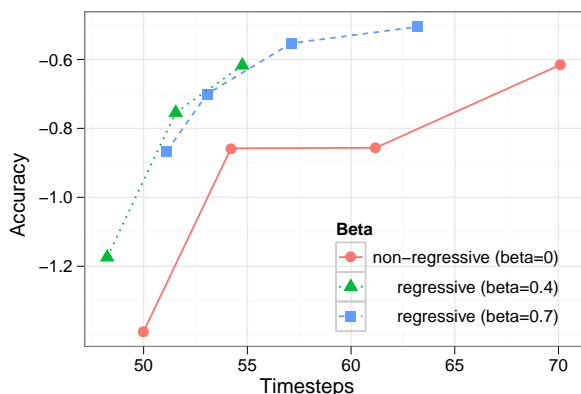


Figure 3: Mean number of timesteps taken to read a sentence and (natural) log probability of the true identity of the sentence ‘Accuracy’ for a range of values of  $\alpha$  and  $\beta$ . Values of  $\alpha$  are not labeled, but increase with the number of timesteps for a constant value of  $\beta$ . For each non-regressive policy ( $\beta = 0$ ), there is a policy with a lower  $\alpha$  and higher  $\beta$  that achieves better accuracy in less time.

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.

## 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 ( $\alpha, \beta$ ) for three different measures of performance, each representing a different trade-off between the importance of accuracy and speed.

### 6.1 Methods

#### 6.1.1 Performance measures

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

$$L(1 - \gamma) - T\gamma \quad (5)$$

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  $\gamma$ . 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.

#### 6.1.2 Optimization of policy parameters

Searching directly for optimal values of  $\alpha$  and  $\beta$  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.<sup>5</sup> Then, we evaluate the model’s performance at each value of  $\alpha$  and  $\beta$  by reading the full test corpus and averaging performance. We then simply use coordinate ascent (in logit space) to find the optimal values of  $\alpha$  and  $\beta$  for each performance measure.

#### 6.1.3 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 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  $\alpha$  and  $\beta$  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.

#### 6.1.4 wFSA implementation

The implementation was the same as in Sim. 1.

#### 6.1.5 Test corpus

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

### 6.2 Results and discussion

The optimal values of  $\alpha$  and  $\beta$  for each  $\gamma \in \{.025, .1, .4\}$  are given in Table 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  $\alpha$  and  $\beta$ , which go up (slower reading) as  $\gamma$  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  $\gamma$  goes down. As predicted, the optimal

<sup>5</sup>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.



| $\gamma$ | $\alpha$ | $\beta$ | Timesteps | Log probability |
|----------|----------|---------|-----------|-----------------|
| .025     | .90      | .99     | 41.2      | -0.02           |
| .1       | .36      | .80     | 25.8      | -0.90           |
| .4       | .18      | .38     | 16.4      | -4.59           |

Table 1: Optimal values of  $\alpha$  and  $\beta$  found for each performance measure  $\gamma$  tested and mean performance at those values, measured in timesteps  $T$  and (natural) log probability  $L$ .

values of  $\beta$  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  $\alpha$  and then return if it falls below  $\beta$ . It would seem, then, that the only reasonable values of  $\beta$  are those that are strictly below  $\alpha$ . 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  $\beta$ ) will generally be higher than when it decided to leave (determined by  $\alpha$ ). 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.

## 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.

## Acknowledgments

The authors thank Jeff Elman, Tom Griffiths, Andy Kehler, Keith Rayner, and Angela Yu for useful discussion about this work. This work benefited from feedback from the audiences at the 2010 LSA and CUNY conferences. The research was partially supported by NIH Training Grant T32-DC000041 from the Center for Research in Language at UC San Diego to K.B., by a research grant from the UC San Diego Academic Senate to R.L., and by NSF grant 0953870 to R.L.

## References

- 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.
- Bicknell, K., & Levy, R. (2010). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In *Proceedings of the 32nd 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), 1–12.
- 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.
- Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109, 193–210.
- 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., & Jones, B. G. (1973). Correlation and letter recognition. *Canadian Journal of Psychology*, 27, 317–326.
- 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.
- 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*, 106, 21086–21090.

- 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 on-line 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).
- Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, *23*, 269–311.
- 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.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.
- 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., 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.
- 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.
- 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.