

# Optimal Processing Times in Reading: a Formal Model and Empirical Investigation

**Nathaniel J. Smith (njs@pobox.com)**

Department of Cognitive Science, 9500 Gilman Drive #515  
La Jolla, CA 92093-0515 USA

**Roger Levy (rlevy@ling.ucsd.edu)**

Department of Linguistics, 9500 Gilman Drive #108  
La Jolla, CA 92093-0108 USA

## Abstract

It is widely known that humans can respond to events they expect more quickly than to unexpected events, but we still have a poor understanding of *why*. Models exist that derive a relation between subjective probability and response time on the basis of optimal perceptual discrimination, but these models rely on the ability of the responder control over perceptual sampling of the environment, rendering them problematic for some domains, such as auditory language processing, in which there are nevertheless clear dependencies between probability and response time. We present a new model deriving the relationship between probability and reaction time as a consequence of *optimal preparation*. This model is valid under very general conditions, requiring only that the results of optimization are invariant across scale of input stimulus granularity. The model makes the strong prediction that response times should scale linearly with the negative conditional log-probability of the stimulus. We present evidence for this prediction in an analysis of an existing database of eye movements in the reading of naturalistic texts.

**Keywords:** Optimal behavior; Language; Response time modeling; Surprisal; Sentence comprehension; Eye movements; Reading

## Introduction

It takes time to perform computation using a physical device, and the human brain is such a device. While obvious, this point is worth revisiting in light of the recent surge of interest in rational models of optimal behavior (Chater et al., 2006; Todorov, 2004). Such models have provided elegant explanations for many aspects of behavior, but processing time provides a particular challenge for this approach. In general, humans respond to different stimuli within any given class with different speeds, and response times are a large part of the stock and trade of experimental cognitive psychology. From an optimality perspective, the difficulty is that it is unclear why the time to spend on performing a computation should ever be larger than the physical minimum. Yet, from a theoretical point of view, response times seem like the perfect candidate for an optimality approach, because they are so clearly relevant to evolutionary fitness. We are all real-time organisms who must react quickly and correctly in a wide variety of circumstances. So why are we still so much slower than we could be?

The primary approach to these problems deployed within the Bayes-optimal framework has been to ascribe reaction times and other such delays to the sensory system. The argument is that we require accurate information about the world to act, but our sensory system is noisy. Therefore, to acquire accurate information, we must wait and gather multiple

noisy samples from which to (Bayes-optimally) average out the noise and extract the signal; if we are optimal perceptual discriminators then the number of samples we must gather (and thus how long we must wait) depends on the form of the signal, of the noise, and of our prior beliefs. (For one example of this approach in the context of word recognition, see Norris, 2006.) It seems unlikely, however, that the sensory system is to blame for *all* response delays; we are reminded every time we start up our computers that computation *qua* computation takes time.

Furthermore, while such models seem plausible for visual perception, it is unclear how they might apply to, for instance, audition. We have reasonable control over how long we look at a scene, but very little control over how long we listen to an utterance. Yet, tasks using single auditorily presented words as stimuli find systematic variation in reaction time — and, in fact, these variations are similar to those observed in corresponding visual presentation paradigms (Goldinger, 1996).

This is one reason that we believe language to be a fruitful area in which to investigate alternative approaches to modeling processing time as an optimal behavior. While language can be presented in the written modality, which is very convenient for experimentation, the bulk of our exposure is to spoken language, and the spoken modality has primacy both evolutionarily and developmentally. To the extent, then, that sensory sampling approaches couched in the framework of optimal perceptual discrimination are implausible for spoken language comprehension, these approaches are unlikely to provide the full story for general language processing. On the other hand, language processing is highly practiced and very efficient, which suggests that some other kind of optimality approach would still be valuable.

In this paper, we present a new model of optimal response time couched in a framework of *optimal preparation* that we believe may be more appropriate to domains such as language processing in which we cannot always control time of sensory exposure. This model is motivated by the well-established fact that processing times in language comprehension are probability-sensitive: in a given context, words which are more predictable are also read more quickly (e.g. Ehrlich & Rayner, 1981). This is intuitively sensible — certainly we would prefer it to the reverse! — but it is, as yet, inadequately theorized. Our model explains this result as optimal behavior under a cost function which trades off preparation costs versus processing time; one would like to pro-

cess quickly, but this requires preparation, and preparation is expensive in its own right. This model makes strong predictions about the relationship between probability, optimal preparation costs, and optimal reading times. We then test the model’s prediction about probability and reading time against a corpus of naturalistic language processing data.

## Model

### How much preparation is too much?

Our main idea is that in general, the nervous system does not operate at the fastest possible speed, and that the reason for this is that operating at the limits of efficiency is very expensive. Instead, it adjusts its performance on particular tasks to optimize a composite cost function that balances the speed achieved against the costs of achieving that speed.

We further assume that this optimization occurs before each stimulus is actually encountered, because once the stimulus is encountered it is too late to reallocate resources — one has whatever resources one has, and all there is to do is to process the stimulus as fast as possible given those constraints. Thus there are two stages, each with a cost: the pre-stimulus or “preparatory” period, where optimization occurs, and the post-stimulus or “processing” period, lasting from stimulus onset until an appropriate response can be made. The duration of the latter is evolutionarily relevant and what we commonly measure experimentally, but the costs incurred in the former are presumably just as important to the brain.

Formally, assume that for a given context there are  $n$  possible stimuli that we may possibly encounter (in the case of reading, these stimuli could be the words that may occur next), and we may freely choose how long each will take us to process, subject to an additive global cost function:

$$C(\mathbf{t}) = \sum_{i=1}^n r(t_i) + E(t_I|\text{context}). \quad (1)$$

Here  $t_i$  is the time we will require to process stimulus  $i$  if it appears,  $\mathbf{t} = \langle t_1, t_2, \dots \rangle$  is the vector of all such times. Our goal is to select  $\mathbf{t}$  in such a way that we minimize our overall cost  $C(\mathbf{t})$ . The cost is composed of two parts. The first term,  $\sum_i r(t_i)$ , corresponds to the preparation cost we incur before encountering the stimulus;  $r(t_i)$  denotes the cost of investing resources to prepare for stimulus  $i$ , and we prepare for all possible stimuli (though, at our option, in differing amounts). The second term,  $E(t_I|\text{context})$ , corresponds to the time it will take to process the stimulus that we do, in fact, encounter; since the stimulus has not yet occurred, its identity is a random variable,  $I$ , and we can only optimize the expected time for processing it. Simple calculus then shows that (1) is minimized when

$$t_i = (r')^{-1}(-p_i) \quad (2)$$

where  $p_i = P(I = i|\text{context})$ , and  $r'$  denotes the derivative. This unwieldy formula becomes clearer if define  $f(x) = (r')^{-1}(-x)$ :

$$t_i = f(p_i). \quad (2')$$

In any case, we are done as soon as we work out what form  $r(t)$  takes. This function summarizes the costs involved in many kinds of preparation occurring over many time-scales. For instance, these might over the short term involve the attentional resources required to speculatively pre-compute responses to stimuli that are especially likely in this context; over the medium term, the metabolic costs of maintaining more or less precise cortical circuit tuning; and over the long term, the allocation of limited cortical area to items which prove themselves to be reliably common. We therefore do not assume or attempt to derive any particular functional form for  $r(t)$  from first principles, and limit ourselves to two simple assumptions: (i) that it depends only on the chosen time  $t_i$  and not on any other specific properties of the context or stimulus; and (ii) that it is some smooth and monotonic decreasing function. We turn instead to the particular attributes of our system of interest, language.

### Scale-free assumption

One of language’s most celebrated properties is that it has hierarchical structure, with regularities occurring at all levels of granularity from, e.g., sentences to clauses to words to morphemes to letters or phonemes. In our experiments we may choose to measure processing time at word granularity, but we have no reason to believe that this is a uniquely preferred scale for the brain. Spoken language is essentially a continuous auditory stream, and clearly there are many options for how to break it into discrete, enumerable ‘stimuli’ as required by our model. Therefore, instead of trying to derive  $r(t)$  directly, let us require that our model give the same answer regardless of the temporal granularity we use to divide our stimuli.

Formally, suppose we have some item  $i$  (e.g., a word) which we can partition into smaller items  $i_1, \dots, i_m$  (e.g., the phonetic segments in that word). Let  $p_{ij}$  denote the conditional probability that  $i_j$  appears given that  $i_1, \dots, i_{j-1}$  have appeared previously — i.e.,  $p_{ij} = P(i_j|i_1, \dots, i_{j-1}, \text{context})$  — while  $p_i$  as defined above can be rewritten as  $P(i_1, \dots, i_m|\text{context})$ . Applying the chain rule to these formulas shows that probability of the larger item is simply the product of the probabilities of the smaller items:  $p_i = \prod_{j=1}^m p_{ij}$ . On the other hand, the time taken to process the larger item is the sum of times taken to process its parts  $t_i = \sum_j t_{ij}$ . By (2'),  $t_{ij} = f(p_{ij})$ . Substituting into (2'), we find

$$\sum_j f(p_{ij}) = f\left(\prod_j p_{ij}\right).$$

That is, the function  $f$  turns products into sums. The only non-trivial functions with this property are logarithms. Working backwards and minding the appropriate monotonicity conditions, we conclude

$$t_i = -\log_k p_i \quad (3)$$

$$r(t) = \frac{k^{-t}}{\log_e k} \quad (4)$$

where  $k > 1$  is a free parameter.

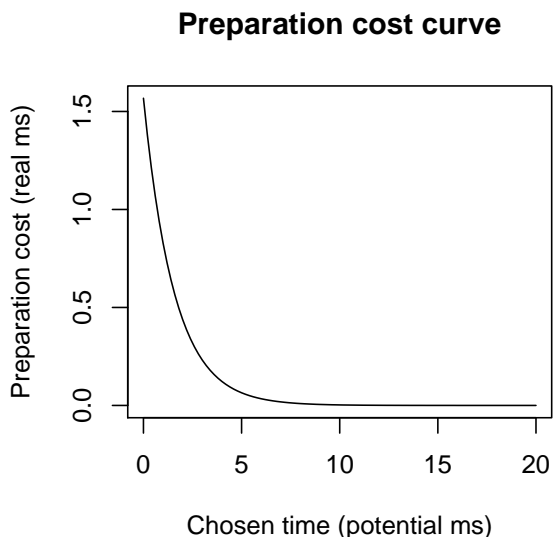


Figure 1: Shape of preparation cost curve  $r(t)$  derived in (4), for  $k = 1.89$  (chosen to match reading time data). X axis gives the different achievable processing times that we select from before encountering the stimulus; Y axis gives the cost for selecting any given time, which in equation (1) is then added to the average actual processing time to produce the total cost.

## Predictions

Our model therefore makes two strong predictions about the processing of stimuli which are continuous and have scale-free hierarchical structure in time: that preparation costs drop off exponentially as the chosen processing time increases (see Figure 1), and that ultimately processing time of a linguistic unit should be proportional to the negative log of that unit’s probability (its *surprisal* or *self-information* in information-theoretic parlance). The latter prediction can be tested using existing data sources.

## Empirical validation

While it is generally agreed that more predictable words are read more quickly, previous work on reading time and probability has suggested many functional forms for this relationship: logarithmic (Hale, 2001; Levy, 2008), linear (Reichle et al., 1998; Engbert et al., 2005), or even reciprocal (Narayanan & Jurafsky, 2004). None have been empirically verified; empirical work has been restricted to factorial comparisons (Rayner & Well, 1996), which provide limited insight into curve shape. In this section, we investigate the relationship directly, using multiple regression techniques on an existing database of eye movements performed during the reading of naturalistic text.

## Methods

**Data** The main technical challenge in measuring the shape of a human response curve is obtaining enough data points to estimate it reliably. Therefore, rather than attempt to construct a small set of balanced stimuli, we chose to analyze the Dundee eye-movement corpus (Kennedy et al., 2003), which consists of all eye-movements made by 10 subjects while reading a collection of newspaper articles totaling approximately 50,000 words. In this paper we report results for first fixation times, a standard reading time measure corresponding to the durations of the first fixation to land on each word in the text.<sup>1</sup> This is not a perfect measure of processing time, and it is not a perfect match to our model (which does not assume that during a fixation centered on some word, subjects will process only that word); these facts will tend to increase the noise in our data. Noise, however, can be overcome through statistical means, and in return we are able to make use of existing methods of word probability estimation, and achieve greater comparability with the existing reading literature.

**Probability estimation** Probabilities were estimated by a trigram language model trained on the 100 million word British National Corpus (BNC). We estimated the model using SRI Language Modeling Toolkit (Stolcke, 2002), with modified Kneser-Ney smoothing (Kneser & Ney, 1995).<sup>2</sup> A trigram model approximates the probability of a word in context  $P(\text{word}_i|\text{context})$  as the probability of the word given two previous words,  $P(\text{word}_i|\text{word}_{i-1}\text{word}_{i-2})$ ; modified Kneser-Ney is a standard method of smoothing these trigram probabilities, and a standard technology for broad-coverage language modeling (Chen & Goodman, 1998). However, it should be noted that Kneser-Ney trigram probability (henceforth, KN3-probability) is still a very noisy estimate of true conditional probability.

**Data selection** The Dundee corpus contains 307,656 first fixations; of these, we eliminated all fixations on words that occurred at the beginning or end of a line, which preceded or followed punctuation, that did not occur in the BNC (i.e., unknown words), or that occurred in the BNC but in segmented form (e.g., the BNC codes *don't* as two words, *do* followed by *n't*). This left  $N = 197,503$  fixations for our analysis, spread roughly evenly across 10 subjects (range: 16666–22390 fixations per subject).

**Confounds** A number of other linguistic measures are correlated with probability, and also known to be correlated with reading time; in particular, these include word length and word frequency. Since we are using naturalistic data, we must control for such confounds retrospectively. Word length is

<sup>1</sup> Analyses of first-pass reading times led to substantially similar results.

<sup>2</sup> Traditionally, such probabilities are estimated via a cloze norming task, but such behavioral measures are impractical for large numbers of data points or low probability events, both of which are major considerations for our data-set.

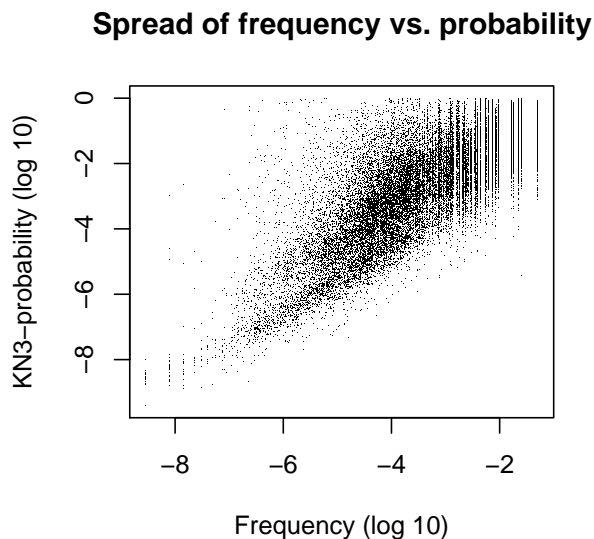


Figure 2: Scatter-plot of log frequency versus estimated log probability for the words in our data.

simply the length of each word in letters. Word frequency is the *unconditional* probability of a word,  $P(\text{word})$ , measured simply as the number of times a word appears in a large body of text, divided by the total number of words in that text. For each fixated word in our corpus, we calculated word frequency from the BNC. Unsurprisingly given their closely related theoretical definitions, frequency is highly correlated with the conditional probability of a word ( $\rho = 0.80$ ), and the log of word frequency is reported to be correlated to processing time on a wide variety of tasks. Our model, of course, suggests that such effects are not driven by frequency *per se*, but rather by probability;<sup>3</sup> however, testing this prediction requires that we analyze our data with respect to both frequency and probability together. Distinguishing such correlated variables relies on what spread does exist; fortunately, this is non-negligible (see Figure 2).

**Analysis** Analysis was carried out in R (R Development Core Team, 2007), using the package `mgcv` for non-parametric multiple regression (Wood, 2006) and `lme4` for mixed-effect linear regression (Bates & Sarkar, 2007).

## Results

**Curve shape** Extracting the shape of an unknown functional relationship from noisy data requires some form of non-parametric regression; doing so while simultaneously controlling for confounds requires multiple non-parametric regression. An elegant framework for such analysis is pro-

<sup>3</sup>Note in particular that in single-word paradigms such as the lexical decision task, there is effectively no context, which means that word probability,  $P(\text{word}|\text{context})$ , and word frequency,  $P(\text{word})$ , give identical predictions.

## Potential relationships between probability and reading time

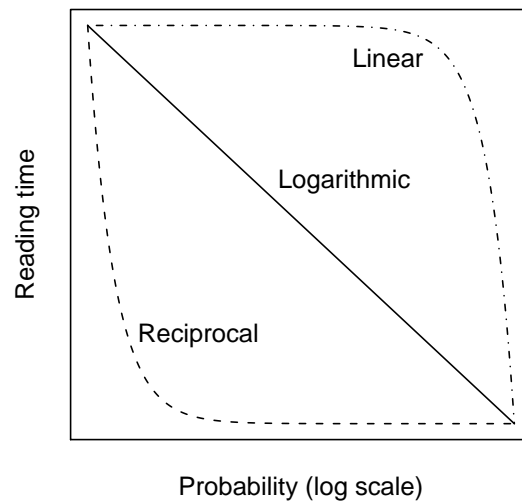


Figure 3: Three possible relationships between probability and reading time as proposed by different authors, illustrated in log-probability space. Our model predicts that the middle curve is correct.

vided by generalized additive models (GAMs), a spline-based extension of the standard linear regression framework (Hastie & Tibshirani, 1990). In our case, we fit a model of the form:<sup>4</sup>

$$\text{Time}_i = \alpha + \beta \cdot \text{WordLength}_i + f(\log \text{KN3-probability}_i) + g(\log \text{Frequency}_i)$$

where  $\alpha$  and  $\beta$  are arbitrary constants, and  $f$  and  $g$  are arbitrary smooth functions; all are chosen by the fitting process. Such an approach, of course, is prone to overfitting; to combat this (and make the problem well-posed), the fit penalizes functions based on how ‘wiggly’ they are, so that there is a trade-off between following the data and avoiding extraneous bends. The relative weight placed on these goals is determined by cross-validation.

The end result of this process are two functions,  $f$  and  $g$  above. Plotting  $f$  will show us how fixation duration varies in response to changes in KN3-probability, after accounting for confounds. Above, three possibilities from the literature were mentioned; the corresponding plots in log-space are shown in schematic form in Figure 3.

The function  $g$  is less immediately relevant, but interesting nonetheless; there is a long tradition of frequency effects in psycholinguistics, but these results are usually confounded with any potential effect of probability (though see Rayner et al., 2004). Examining  $g$  will show us the residual effect of frequency after accounting for the effects of probability.

<sup>4</sup>In this model, logs are taken of KN3-probability and frequency purely for convenience; invertible transformations of predictor variables have a minimal effect on non-parametric regression.

One fit was performed to the data from each subject individually, with results illustrated in Figure 4. As predicted by our model, the curves in the left column (KN3-probability) are very close to linear (in log space), and this pattern holds across at least five orders of magnitude in probability. Nine out of ten subjects show this pattern; the exception is subject G, whose effect appears to be minimal, if any. The curves on the right (frequency) are also roughly linear, and of a similar order of magnitude, though they appear to be less reliable — only seven out of ten subjects show a clear effect.

Now that we have established the shape of these effects, we can better quantify their strength and significance using traditional parametric techniques.

**Significance** To analyze significance, we used linear regression of fixation duration on word length, log frequency, and log KN3-probability, with subject as a random effect.

After controlling for word length and frequency, KN3-probability remains highly significant ( $\chi^2(2) = 246.12, p \ll 0.001$ ) as a predictor of first-fixation times. After controlling for word length and KN3-probability, frequency also remains significant ( $\chi^2(2) = 182.81, p \ll 0.001$ ).

We can also investigate the relative magnitude and reliability of these effects by fitting a model including both frequency and KN3-probability simultaneously, after standardizing them both to ensure comparability. In such a model, the response coefficient for KN3-probability is both larger than that for frequency ( $-4.1$  vs.  $-3.7$  in arbitrary units), and less variable across subjects (standard deviation  $1.5$  vs.  $2.4$ ).

### Discussion

We have presented a model which predicts that in language-like tasks — those where processing is skilled, speed is important, and stimuli are arranged through time in a continuous manner with no preferred scale — the processing time for an individual unit should be proportional to the negative log-probability of that unit occurring. Further, we have for the first time performed a broad-coverage analysis of the functional relationship between probability and reading times, and have found that over a wide range of probabilities, this effect is both significant and takes the predicted form. Although we have validated this prediction of the model on language processing, this model could in principle apply to any cognitive or perceptual domain in which our assumption that there is no preferred granularity scale of processing is reasonable.

These results suggest that a substantial portion of what have previously been understood as frequency may, in fact, be context-sensitive probability effects — which may be problematic for theories that explain log-frequency effects as a result of the (static, context-insensitive) structure of the lexicon, such as the classic logogen theory of (Morton, 1969) and its descendants in this respect, such as READER (Just & Carpenter, 1992). It still remains to compare to the other covariates that have been proposed besides word frequency (e.g., Gernsbacher, 1984; Morrison & Ellis, 1995; McDonald & Shillcock, 2001; Murray & Forster, 2004).

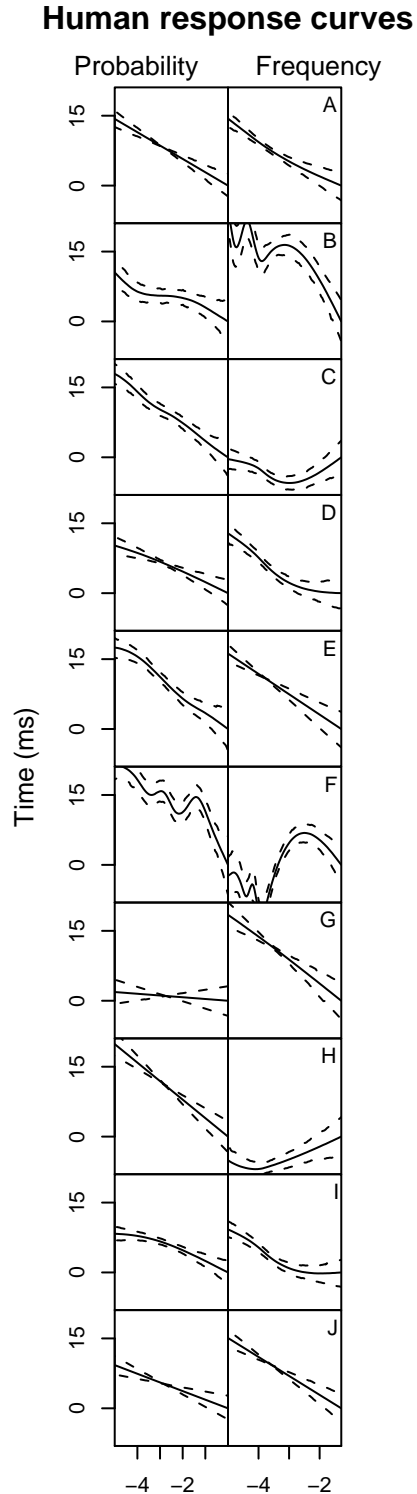


Figure 4: Estimated response curves for probability (left column) and frequency (right column), both in log space, plotted individually by subject. The X axis in each case ranges from  $-5$  to the maximum value that occurred in the data set. This range was chosen to include  $\approx 90\%$  of all data points (see Figure 2); outside of this range the fit becomes extremely unreliable.  $X = -5$  corresponds to a 1-in-100,000 event. Dashed lines are bootstrapped 95% confidence intervals.

On the other hand, frequency does remain significant, and it is not entirely clear how to interpret this. (No extant theories have proposed explanations for why we would expect probability and frequency to have separate, independent effects.) One possibility is that this result arises from noise in the probability estimation process: trigram models have far more parameters than unigram (word frequency) models, which increases estimation error; furthermore, the theoretical quantity of interest is not trigram probability at all, but ‘full’ conditional probability, and this approximation introduces additional error. Taken together, this means that our probability estimates should be understood as having a much higher degree of noise than our frequency estimates. Despite this noise, however, KN3-probability still marginally outperforms frequency in explaining fixation times. This suggests the possibility that frequency’s role could diminish or conceivably disappear with improvements in our technical ability to estimate subjective probability.

However, it is also possible that frequency’s role will not disappear, and is in fact real. Our model makes the strong claim that subjective probability is the only determinant of reading time, so if frequency’s role is validated, there are two possibilities: either our model is essentially correct but humans are sub-optimal with regards to estimating conditional probabilities of words in text — perhaps their estimates are partially biased and smoothed by word frequency, as a quick, dirty, and low-variance approximation — or our model is incomplete.

Finally, we should note that the juxtaposition of our optimal-preparation model against optimal perceptual discrimination models such as Norris’s (2006) Bayesian Reader opens up a typology of optimal response time theories.<sup>5</sup> Optimal-discrimination and optimal-preparation accounts make different predictions — notably, the perceptual confusability of the stimulus should have a huge effect on response times in optimal-discrimination accounts, whereas it does not play a role in our account. It is also logically possible that the truth lies in a combination of both accounts, or in a third account that lies elsewhere in this typology, in a theory of optimal response times yet to be constructed.

## Acknowledgments

This work was supported by an NSF graduate fellowship to NS.

## References

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using Eigen and Eigen++ [Computer software]. (R package version 0.99875-9.1)  
*The British National Corpus, version 3 (BNC XML edition)*. (2007). (Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: <http://www.natcorp.ox.ac.uk/>)

<sup>5</sup>Norris’s Bayesian Reader also predicts a response-time effect linear in negative log-probability, for reasons explained in (Norris, submitted).

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10, 287–291.

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

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6), 641–655.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation in reading. *Psychological Review*, 112(4), 777–813.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113, 256–281.

Goldinger, S. D. (1996). Auditory lexical decision. *Language and Cognitive Processes*, 11(6), 559–567.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAAACL-2001* (pp. 159–166).

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122–149.

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

Kneser, R., & Ney, H. (1995). Improved backing-off for *m*-gram language modeling. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (Vol. 1, pp. 181–184).

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

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3), 295–323.

Morrison, C., & Ellis, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 116–133.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76(2), 165–178.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, 111(3), 721–756.

Narayanan, S., & Jurafsky, D. (2004, November). *A Bayesian model of human sentence processing*. (Unpublished manuscript, <http://www.icsi.berkeley.edu/~snarayan/newcog.pdf>)

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113(2), 327–357.

Norris, D. (submitted). *Why the lexical decision task really does tell us a lot about word recognition: Modeling reaction-time distributions with the Bayesian Reader*.

R Development Core Team. (2007). *R: A language and environment for statistical computing* [Computer software and manual]. Vienna, Austria. (ISBN 3-900051-07-0)

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(4), 720–732.

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

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, 105(1), 125–157.

Stolcke, A. (2002). SRILM—an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing* (Vol. 2, pp. 901–904).

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, 7(9), 907–915.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton: Chapman and Hall/CRC.