A rational model of word skipping in reading: ideal integration of visual and linguistic information

Yunyan Duan (yduan@u.northwestern.edu)
Department of Linguistics, Northwestern University
Evanston, IL 60201 USA

Klinton Bicknell (klinton@duolingo.com)
Duolingo AI Research
Pittsburgh, PA 15206 USA

Northwestern University
Evanston, IL 60201 USA

Abstract

During reading, readers intentionally do not fixate a word when highly confident in its identity. In a rational model of reading, word skipping decisions should be complex functions of the particular word, linguistic context, and visual information available. In contrast, simple heuristic of reading only predicts additive effects of word and context features. Here we test these predictions by implementing a rational model with Bayesian inference, and predicting human skipping with the entropy of this model's posterior distribution. Results showed a significant effect of the entropy in predicting skipping above a strong baseline model including word and context features. This pattern held for entropy measures from rational models with a frequency prior but not from ones with a 5-gram prior. These results suggest complex interactions between visual input and linguistic knowledge as predicted by the rational model of reading, and a dominant role of frequency in making skipping decisions.

Keywords: eye movements; reading; word identification; rational analysis; skipping

Introduction

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

One promising model of this type comes from a perspective of rational analysis. The idea is to consider the reading process as one that combines information from various sources to identify words and then makes eye movement decisions to maximize identification efficiency (Bicknell & Levy, 2010, 2012; Legge, Klitz, & Tjan, 1997; Legge, Hooven, Klitz, Mansfield, & Tjan, 2002). In previous rational models of reading, text identification process is modeled using Bayesian inference that combines two sources of information: (1) probabilistic knowledge of the structure of the language, serving as the prior, and (2) uncertain visual evidence, serving as the likelihood. Given a prior and a particular set of visual evidence, probabilistic inference yields a posterior distribution on the text, which specifies the probability of each possible identity of the text. The role of eye movements in this analysis is to obtain particular pieces of visual evidence, and the most efficient, rational reading behavior will be to use the current posterior distribution to determine the most useful time and place to move the eyes next. Therefore, any eye movement behaviors explained by this model of reading can be seen as naturally born from one simple origin: the rational gathering of visual evidence for text identification.

In contrast, the dominant models of eye movement control in reading tend to use heuristic linking hypothesis between text identification and eye movements (e.g., E-Z Reader, Reichle et al., 2009; and SWIFT, Engbert et al., 2005). In these models, eye movements are driven by a word identification process that is represented with discrete states (e.g., not identified, partially identified that leads to saccade programming, fully identified), the transitions between which depend on a certain amount of durations computed from a few coarse visual and linguistic variables of the word. For example, in E-Z Reader the duration of L1 and L2 depend on a stochastic function of two linguistic variables, the word's frequency in the language and its predictability in context, and one visual variable, the average distance from each of its letters to the point of fixation. After spending the pre-computed duration needed to achieve a certain stage of word identification to begin programming a saccade and then achieve complete identification of the current word, the model moves eyes to (roughly) the center of the next word to be identified. The role of eye movements in this heuristic model is a direct reflection (with stochastic noise) of cognitive process identifying a word, the difficulty of which depends on coarse properties of the word as a whole.

There are situations where word identification can be completed with more fine-grained knowledge about the particular word than merely coarse information, and we would like to make precise predictions about eye movement behaviors accordingly. Consider situations where visual information about only the beginning of some words is enough for identification, e.g., seeing the initial letters 'xyl' of the word 'xylophone' (Hyönä, Niemi, & Underwood, 1989). Similarly, in certain linguistic contexts, a reader only needs to see a few of

the initial letters of a word to be confident in its identification, such as in 'The children went outside to pl…'. Do readers in fact combine more fine-grained information than simply word frequency and word length in the way as predicted by rational models of reading?

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

Previous empirical research finds that readers' likelihood of skipping a word increases with short word length, close launch sites to the word, high word frequency, and high contextual predictability (Rayner, 1998). Regarding how different sources of information may interact in skipping, studies of skipping short words and especially the word *the* suggest that visual information and word frequency information trump information from the sentence context (Angele & Rayner, 2013; Angele, Laishley, Rayner, & Liversedge, 2014). Despite these findings, the fine-grained predictions of a rational model can be better tested with a set of eye movement decisions that happen in natural reading and that have wide variation in visual and linguistic information available to the reader. The goal of the current paper is to directly test the fine-grained predictions using word skipping, and to gain insights into the role of different sources of information in making skipping decisions.

Related work

Empirical findings about skipping

At the aggregate level, the effects of visual and linguistic variables on skipping are very robust. Word length is considered to play a more important role than any other factors, as found in a meta-analysis showing that word length explained more variance than word frequency and predictability in regression models predicting skipping rate (Brysbaert, Drieghe, & Vitu, 2005). The effect that close launch sites increase skipping rates is also strong and robust (Brysbaert et al., 2005). As for linguistic variables, there is abundant experimental evidence that skipping rate increases as word frequency increases (Rayner, Sereno, & Raney, 1996; Angele et al., 2014), and that high predictability leads to high skipping rate (Balota, Pollatsek, & Rayner, 1985; Rayner, Slattery, Drieghe, & Liversedge, 2011). Predictability is usually measured as cloze probability, varying across conditions either with different sentential frames or target words (Balota et al., 1985; Rayner et al., 2011). The effects hold in corpus analysis as well, as Luke and Christianson (2016) find that high target predictability lead to more word skipping for both content and function words. Kliegl, Grabner, Rolfs, and Engbert (2004) also find significant effect of predictability, word length, and word frequency on skipping rate using regression analyses on Potsdam Sentence Corpus, though not including any interactions among these factors.

Several studies have looked into the interactions between visual and linguistic factors on a coarse level. One approach is to analyze linguistic effects on data split in launch sites in post-hoc analysis. For example, Rayner et al. (1996) observe reliable frequency effect on skipping rate at near launch sites (> -5) but not at far launch sites, and White, Rayner, and Liversedge (2005) find significant interaction between predictability and word length preview overall, which diminish to non-effect for far launch sites (near launch sites are defined as those ≥ -3, while far launch sites are those ≤ -4). Another approach to study the interaction of visual and linguistic information is to manipulate parafoveal preview. A preview of the definite article *the* increases readers' skipping rate, even when syntactic constraints do not allow for articles to occur in that position (Angele & Rayner, 2013; Angele et al., 2014). Skipping rates are higher for the preview of a highly predictable word or its visually similar nonword counterpart than the preview of a low-predictability word (Balota et al., 1985), and for the preview of a predictable word than for a visually similar nonword (Drieghe, Rayner, & Pollatsek, 2005). Staub and Goddard (2019) observe that frequency effect on skipping rate is maintained with both valid and invalid preview, but predictability influences skipping only with valid preview. Additionally, English readers only benefit from the preview of a semantically similar neighbor in highly-constraining context but not in moderate-constraining context (Schotter, Lee, Reiderman, & Rayner, 2015).

In sum, previous research have identified visual and linguistic factors that influence skipping by conducting reading

experiments and corpus studies. There is also evidence for interactions between visual and linguistic factors, but they are constrained to a small set of well-controlled language materials and analyzed on a coarse level. A systematic analysis with skipping made for a variety of words in a variety of contexts with a variety of launch sites would help gain insights into how visual and linguistic variables interacts to identify a word before fixating it and skip at a fine-grained level.

Other instances of rational models of reading

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

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

Rational model of skipping

Word identification as Bayesian inference

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

$$p(w|\mathcal{I}) \propto p(w)p(\mathcal{I}|w) \tag{1}$$

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

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

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

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

Single word belief updating

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

$$
\begin{aligned}
\mathbf{x}_i^{(t)} &= \log \frac{p(w_i|\mathcal{I}^{(0,\ldots,t)})}{p(w_v|\mathcal{I}^{(0,\ldots,t)})} \\
&= \log \frac{p(\mathcal{I}^{(t)}|w_i)p(w_i|\mathcal{I}^{(0,\ldots,t-1)})}{p(\mathcal{I}^{(t)}|w_v)p(w_v|\mathcal{I}^{(0,\ldots,t-1)})} \\
&= \log \frac{p(\mathcal{I}^{(t)}|w_i)}{p(\mathcal{I}^{(t)}|w_v)} + \log \frac{p(w_i|\mathcal{I}^{(0,\ldots,t-1)})}{p(w_v|\mathcal{I}^{(0,\ldots,t-1)})} \\
&= \Delta\mathbf{x}_i^{(t)} + \mathbf{x}_i^{(t-1)}
\end{aligned} \tag{3}
$$

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

Experiment

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

Baseline model

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

Table 1: Generalized additive mixed-effects regression model results of baseline model (note that random slopes for these fixed effects were not included in the model; the model included a random intercept over participants). The GAMM was fitted by REML, and p-values were reported using *summary.gam* function in *mgcv* package (Wood, 2011).

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

Model We analyzed first-pass skipping in the Dundee corpus with a generalized additive mixed-effects regression model (GAMM) predicting skipping from a wide range of variables previously shown to influence skipping, including word length, launch site, word frequency, surprisal, and contextual constraint. We estimated word frequency (log unigram probability) and 5-gram surprisal (log 5-gram probability) with n-gram models (Goodkind & Bicknell, 2018) trained on Google One Billion Word Benchmark (Chelba et al., 2013), and we measured contextual constraint as the entropy of the 5-gram probability distribution of words in a vocabulary of 20,001 words. We defined the vocabulary to include all words that were in both the Dundee corpus and our language modeling corpus, plus words with frequencies above a cutoff chosen such that the resulting total vocabulary would have about 20,000 words. We also controlled for the previous word's properties such as word length and frequency, and included a random intercept over participants. Crucially, this GAMM allowed for non-linear effects of each of these variables, providing a strong baseline. Table 1 shows all the fixed effects in the baseline model.

Rational model

Simulation For each observation in the dataset, we simulated 50 trials using the rational model of skipping for each parametrization of the model. In each trial, a piece of visual information from the launch site is sampled and combined with the linguistic information to generate a posterior distribution of possible identities of the word. As described above, the visual information in this model has two param-

Table 2: Significance of averaged entropy of a rational model's posterior distribution when added to the baseline model.

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

eters: overall visual input quality Λ and the width of acuity function σ. We used fifteen sets of parameter pairs for the models; these parameters were chosen to be values that spanned a wide part of the parameter space while also respecting the trade-off between width of the acuity function and its overall quality.[1] The linguistic information (prior) in this model is given by either the word frequency (unigram) or 5-gram language models, as used in our baseline model.

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

Results

Baseline model

GAMM results of the baseline model are summarized in Table 1. The results confirm previous findings that word length, launch site, frequency, surprisal, and contextual constraint significantly influenced human skipping. Moreover, this baseline model captures non-linear interactions among these predictors, indicating that different sources of information interactively guide skipping at an aggregated level.

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

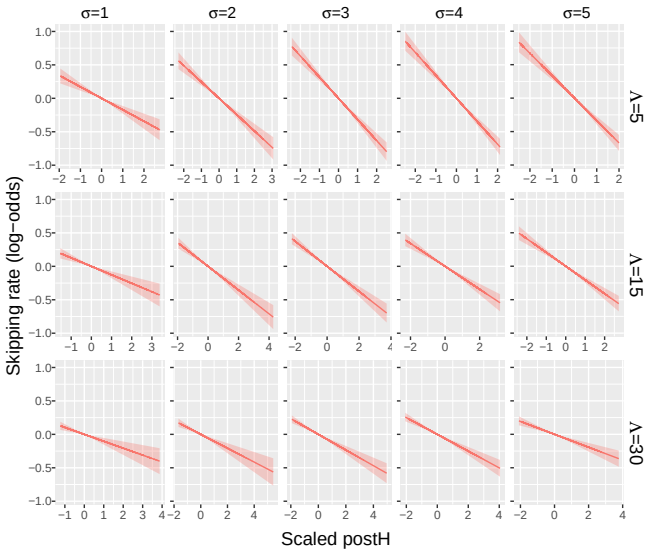

Figure 1: Partial effect of postH with a frequency prior in predicting skipping rate.

Rational model

The partial effects of postH computed from the GAMMs are visualized in Figure 1 (frequency prior) and Figure 2 (5-gram prior), after controlling for all variables in the baseline model and additionally a random slope of postH over participants. The significance of postH when added to the baseline model is reported in Table 2. For postH from rational models with a frequency prior, the effects are significant in the predicted direction: high postH indicates high uncertainty about the word's identification and is associated with lower skipping rates; these effects are robust to parameter choice and are significant for all parametrizations tested. For postH from rational models with a 5-gram prior, the effects are generally not significant, though they do all trend in the same direction and show the pattern that skipping rates increase as the uncertainty over the word's identity increases, opposite to the predicted direction.

Discussion

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

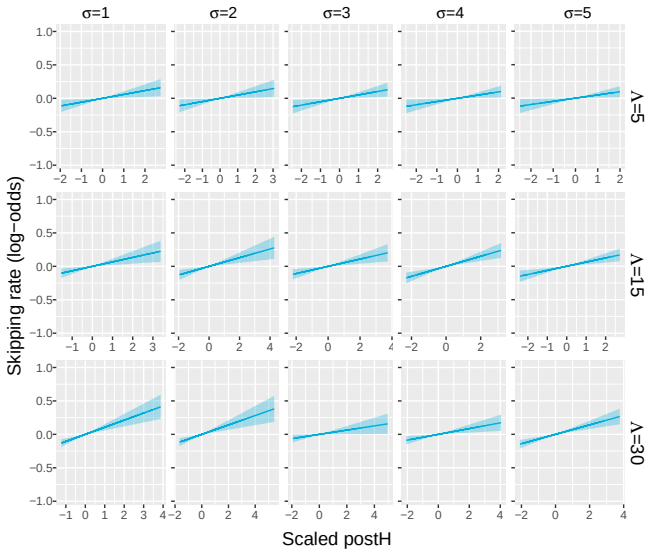

Figure 2: Partial effect of postH with a 5-gram prior in predicting skipping rate.

rate, while the trend of the effect of postH from models with a 5-gram prior is in an opposite direction.

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

The observation that postH from a frequency prior better predicts skipping than the 5-gram prior is potentially problematic for a fully rational model of skipping, though: a reader that maximize usage of all the information available should be better predicted by a model with 5-gram prior than one with frequency prior. Rather, this pattern lines up with previous findings on the skipping of *the*, which relies on visual and frequency information more than structural information (Angele et al., 2014). This pattern is also consistent with the finding that frequency effect but not predictability effect on skipping survives bad parafoveal visual input, which may be explained by different time course of frequency and contextual information in making eye movement decisions (Staub & Goddard, 2019). A possible reason of our finding is that skipping decisions may be made without full knowledge of the context, leading to the absence of effect from our measure (i.e. 5-gram) of contextual information. Specifically, since saccade programming takes a relatively long time and identification/processing of the fixated word continues during this lag, skipping decisions may be made before the previous word is fully identified and integrated into the context. In spite of this issue to be further examined, we find that the entropy of a posterior distribution from a frequency prior improves prediction of skipping with average variables, suggesting a complex combination of information sources as predicted by rational models of reading.

Acknowledgments

This research was supported by the National Science Foundation under NSF 1734217.

References

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

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

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive psychology*, *17*(3), 364–390.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 1168–1178).

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

Brysbaert, M., Drieghe, D., & Vitu, F. (2005). Word skipping: Implications for theories of eye movement control in reading. *Cognitive processes in eye guidance*, 53–77.

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

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

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

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

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

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

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

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153–168.

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

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. chips 2002: New insights from an ideal-observer model of reading. *Vision Research*, *42*(18), 2219–2234.

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

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive psychology*, *88*, 22.

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

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

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 514.

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

Schotter, E. R., Lee, M., Reiderman, M., & Rayner, K. (2015). The effect of contextual constraint on parafoveal processing in reading. *Journal of memory and language*, *83*, 118–139.

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

White, S. J., Rayner, K., & Liversedge, S. P. (2005). The influence of parafoveal word length and contextual constraint on fixation durations and word skipping in reading. *Psychonomic bulletin & review*, *12*(3), 466–471.

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