Research Report

EYE MOVEMENTS REVEAL THE ON-LINE COMPUTATION OF LEXICAL PROBABILITIES DURING READING

Scott A. McDonald and Richard C. Shillcock

University of Edinburgh, Edinburgh, Scotland, United Kingdom

Abstract—*Skilled readers are able to derive meaning from a stream of visual input with remarkable efficiency. In this article, we present the first evidence that statistical information latent in the linguistic environment can contribute to an account of reading behavior. In two eye-tracking studies, we demonstrate that the transitional probabilities between words have a measurable influence on fixation durations, and using a simple Bayesian statistical model, we show that lexical probabilities derived by combining transitional probability with the prior probability of a word's occurrence provide the most parsimonious account of the eye movement data. We suggest that the brain is able to draw upon statistical information in order to rapidly estimate the lexical probabilities of upcoming words: a computationally inexpensive mechanism that may underlie proficient reading.*

The exploitation of redundancy in the input is a pervasive processing strategy; research in areas of visual perception such as motion detection (e.g., Krauzlis & Adler, 2001) suggests that the brain's ability to anticipate what is to come next could also be deployed during reading. One potential source of information that could assist prediction[1] is the statistical knowledge implicit in readers' input: word-to-word contingency statistics, or transitional probabilities. Within a language, there are words that have a high probability of following a given word (e.g., *on* often comes after *rely*), entailing that the occurrence of one word can be confidently predicted from the occurrence of the other. The relevance of statistical information of this kind for essential components of language development (Saffran, Aslin, & Newport, 1996) and its influence on the phonological characteristics of adult language production (Jurafsky, Bell, Gregory, & Raymond, 2001) have already been demonstrated.

We asked two questions addressing the potential connection between statistical information latent in the linguistic environment and reading behavior. First, does statistical knowledge in the form of transitional probabilities influence the relative ease or difficulty of lexical processing in reading? Second, what is the relative importance of context-dependent and context-independent statistical information? The higher a word's frequency of use, the less time the eyes spend on it (e.g., Just & Carpenter, 1980; Rayner, 1998; Rayner, Sereno, & Raney, 1996). Because transitional probability and frequency covary, is it a word's predictability or its a priori probability of occurrence that is the more relevant variable? Alternatively, is the best account provided by a probabilistic view that integrates the two measures?

Address correspondence to Scott McDonald, Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, Scotland EH8 9JZ; e-mail: scott.mcdonald@ed.ac.uk.

1. We do not mean "prediction" in any sense of being explicit, conscious, or strategic.

EXPERIMENT 1

In the first of two experiments, participants were required to read sentences containing contiguous verb-noun sequences that varied in their transitional probability. Sentence pairs were constructed such that the length and corpus frequency of the nouns were closely matched, and the neutral prior context was held constant (e.g., high probability—*One way to avoid confusion is to make the changes during vacation*; low probability—*One way to avoid discovery is to make the changes during vacation*). Only the transitional probability of the verb-noun pair was varied. The high- and low-probability sentences were matched for their rated plausibility by a separate group of participants.

Method

The starting point for the materials was a set of 48 verbs. Each verb was paired with a highly predictable and a less predictable (though still plausible) noun object, based on transitional probabilities computed from the 100-million-word British National Corpus (Burnage & Dunlop, 1992). Transitional probabilities were estimated as follows: $P(\text{noun}|\text{verb}) = [\text{frequency}(\text{verb},\text{noun})/\text{frequency}(\text{verb})]$; the mean values were .01011 and .00038 for the high- and low-predictability sentences, respectively. Each item was closely controlled for the length and corpus frequency of the noun. Two sentences were constructed for each item, with an identical neutral context preceding the critical verb-noun pair.

Sentences were rated for plausibility by an independent group of 26 participants from the same population as the participants in the eye-tracking experiment. Ratings were made on a 7-point scale. There was no difference in mean plausibility between the high- and low-probability sentences (high: $M = 5.3$, low: $M = 5.2$), $t(47) = 0.62$, $p = .54$. The materials were also assessed for predictability using the Cloze procedure (Taylor, 1953). Sixteen additional participants were presented with each sentence up to and including the verb, and were asked to supply the first word that came to mind that could plausibly continue the sentence. Although the percentage of participants producing the selected noun was larger for the high-probability than the low-probability sentences (7.96% compared with 0.79%), this difference is much smaller than the Cloze manipulation typically required in order to observe a predictability effect (cf. Rayner & Well, 1996).[2]

We employed a repeated measures design, creating two versions of the materials; each version contained 24 high- and 24 low-probability sentences, which were interspersed with 48 other sentences of similar structure: 24 from a separate experiment and 24 fillers. Each filler sentence was followed by an untimed yes/no comprehension question.

Twenty-four young adults were each paid £5 to take part; all were native English speakers and had normal or corrected-to-normal (with soft contact lenses) vision. Participants were seated at a viewing dis-

2. A full list of the stimulus materials with their associated transitional probabilities, rated plausibilities, and Cloze values is available on the Web at http://www.iccs.informatics.ed.ac.uk/~scottm/lexical_probability.html.

648   VOL. 14, NO. 6, NOVEMBER 2003

tance of 75 cm from a 15-in. RM VGA monitor. Stimuli were displayed in a monospaced font as white letters on a black background, and occupied a single line of the display. One degree of visual angle was equivalent to 3.8 characters. Eye movements were recorded from the right eye using a Fourward Technologies Generation 6.3 Dual Purkinje Image eyetracker (resolution of less than one min of arc), which was interfaced to a 486 personal computer. Gaze position was sampled every millisecond. A forehead rest and bite bar were employed to minimize head motion. Calibration of the eyetracker was checked and adjusted periodically throughout the course of the experiment.

Results and Discussion

The data for 3% of the critical trials were lost because of blinks or track losses. Abnormally long fixations (> 700 ms) were also excluded from analysis. Table 1 shows the effect of predictability on several early eye movement measures. The principal result involves the duration of participants' initial fixation on the target nouns. Initial-fixation duration is a measure of processing effort that is sensitive to variables such as a word's frequency of use (Just & Carpenter, 1980; Rayner, 1998; Rayner et al., 1996) and its predictability from context (Balota, Pollatsek, & Rayner, 1985; Ehrlich & Rayner, 1981; Rayner & Well, 1996). Analyses of variance, treating either participants (F_1) or items (F_2) as a random factor, showed that this duration was shorter for verb-noun combinations with a high transitional probability than for pairs with a low transitional probability.

Figure 1 (left panel) displays the relationship between initial-fixation duration and launch distance (the distance in character spaces between the previous fixation and the beginning of the target word). The processing advantage for the more predictable nouns tended to increase as launch distance decreased, Spearman $\rho = -.69$, $p = .058$. This trend can be explained by the fact that parafoveal preview of the target noun is more viable the closer the previous fixation. Parafoveal preview increases the efficiency of lexical processing, as demonstrated by studies of reading behavior when no preview is available (Balota et al., 1985; Inhoff & Rayner, 1986). It appears that in order for statistical information about the probability of the target noun following the

verb to come into play, at least partial visual information about the noun needs to be available during the processing of the verb. Parafoveal visual information and statistical information may facilitate lexical processing by converging on a specific lexical representation.

There is substantial evidence that a word's predictability in its sentential context has a clear influence on the ease with which it is processed (e.g., Balota et al., 1985; Kutas & Hillyard, 1984; Rayner & Well, 1996; Zola, 1984). However, effects of predictability may be due to high-level knowledge, in which the meaning derived from integrating the meanings of the individual words in the previous context with knowledge about the world forms the basis of expectations about upcoming words (information conceivably contributing to Cloze test responses), or to information of the sort that would be provided by transitional probabilities. Our experiment is the first attempt to disentangle these two sources of predictability, and our findings indicate a unique contribution of statistical information to reading behavior.[3]

Figure 1 (right panel) displays the distributions of first-fixation durations for the high- and low-probability conditions. Transitional-probability effects emerged at approximately 150 ms, closely comparable to the 130 to 175 ms reported for the emergence of word frequency effects (Sereno, Rayner, & Posner, 1998; Vitu, McConkie, Kerr, & O'Regan, 2001). It has been claimed that because frequency and predictability fail to interact in factorial experiments, they affect distinct processing stages during reading (Altarriba, Kroll, Sholl, & Rayner, 1996; Inhoff, 1984). Our eye-tracking evidence suggests instead that they may influence the same stage of reading, as the effects of frequency and statistically defined lexical predictability have a similar locus in the time course of lexical processing.

One potential deficiency of transitional probability as a measure of lexical predictability is that it is estimated using the relative frequency from a corpus, and so does not take into consideration the amount of evidence underlying the value. For example, the transitional probability $P(\text{word}_2|\text{word}_1) = .3$ can result from word_2 co-occurring with word_1 three times (if word_1 has a frequency of 10) or 300 times (if word_1 has a frequency of 1,000); clearly, the latter case provides a more reliable estimate for the value of word_1 as contextual evidence for word_2. A second issue concerns the relative importance of contextual evidence versus context-free predictions about a word's occurrence. Although the transitional probability $P(\text{havoc}|\text{wreak})$ is high, because *wreak* is nearly always followed by *havoc* in text, the a priori probability of *havoc* is very low. Lexical predictability may ideally reflect an integration of both types of probabilistic information.

Bayes' law provides a principled approach for weighting and combining the evidence for the outcome of an event with prior information, in order to compute the likelihood of observing the event given the evidence; for our purposes, this translates to estimating lexical probabilities. We considered a word's corpus frequency to represent prior knowledge about its occurrence, and modeled a word's occurrence in context as a binomially distributed random variable, where its relative frequency is interpreted as the number of "successes." For example, if

Table 1. *Eye movement measures for the target nouns in Experiment 1*

Measure	Transitional probability		Comparison between conditions			
	High	Low	$F_1(1, 23)$	p	$F_2(1, 47)$	p
Initial-fixation duration	261	272	4.88	.037	4.30	.044
Gaze duration	291	303	3.01	.096	2.81	.100
Single-fixation duration	261	274	7.08	.014	3.75	.059
Probability of skipping	.114	.095	1.33	.261	0.91	.346

Note. Fixation durations are in milliseconds and represent the mean value for each participant averaged across participants. Gaze duration is the summed duration of all fixations made on a word during first-pass reading; single-fixation duration is the fixation time on words receiving only one fixation.

3. Support for high-level knowledge driving our results would be provided if the largest Cloze value differences corresponded to the largest duration differences. However, when we removed the items ($n = 6$) whose high- and low-transitional probability nouns differed in Cloze by more than 25%—which brought the mean Cloze value for the high-probability condition down to 3.59%—we obtained nearly identical results.
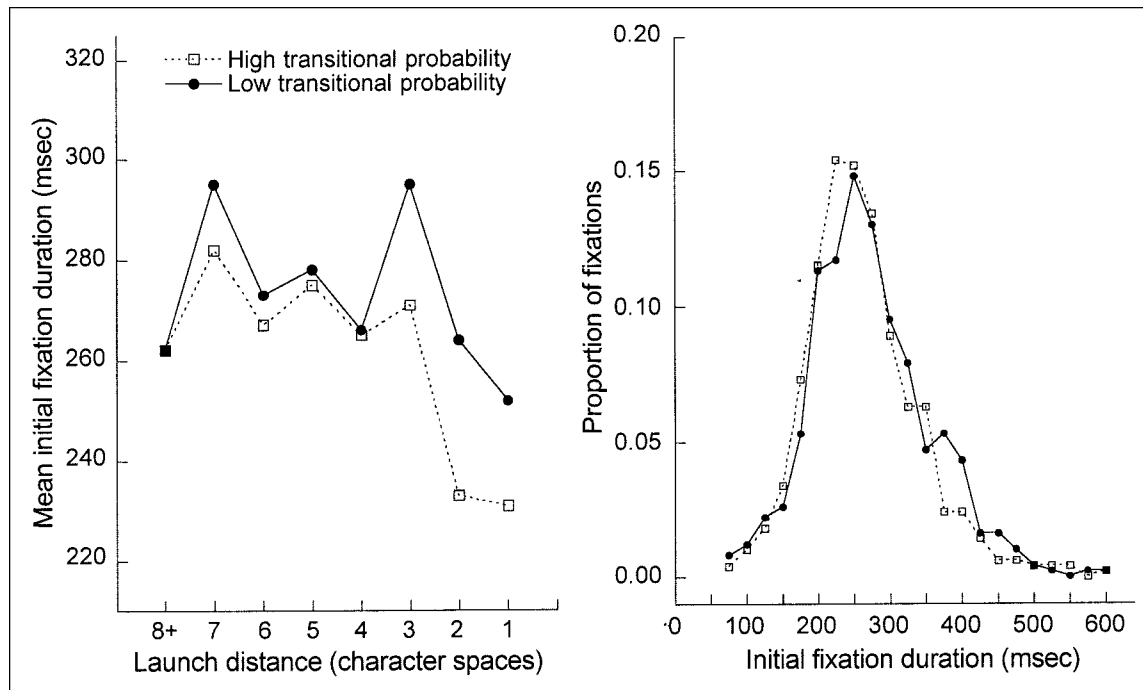
Fig. 1. Effect of transitional probability on the duration of initial fixations in Experiment 1. The left panel shows mean first-fixation duration as a function of predictability and launch distance (the number of character spaces between the last fixation and the space before the target word). The right panel shows the frequency distributions of initial-fixation durations for target words with high and low transitional probabilities. Bin size is 25 ms.

wreak occurs 100 times in a corpus, and the relative frequency of *havoc* following *wreak* is 70/100, then the number of successes, $s_{evidence}$, is 70 and the number of failures, $f_{evidence}$, is 30. By assuming a beta prior distribution (the conjugate prior for the binomial), the expected value of the posterior density can be easily computed:

$$E[P(\text{word}_2|\text{word}_1)] = \frac{\alpha \cdot s_{prior} + s_{evidence} + 1}{\alpha \cdot f_{prior} + f_{evidence} + 2}$$

In this equation, s_{prior} is simply the corpus frequency of word_2, and f_{prior} is defined as $N - \text{frequency}(\text{word}_2)$, where N is the corpus size in words. The weighting parameter α encodes the relative importance of prior knowledge and contextual evidence to the posterior probability. We set this parameter empirically, by finding the optimal linear fit to the data for first-fixation duration collected from the filler sentences in Experiment 1. The best fit was obtained with $\alpha = .0001$. Note that this parameter itself could be allowed to vary, subject to a density function that captures the dependence of α on other variables, and certainly more sophisticated models of word occurrence than the binomial distribution could be employed. Our approach represents a reasonable starting point.

EXPERIMENT 2

The aim of Experiment 2 was to assess the relative importance of the frequency, transitional probability, and Bayesian posterior probability measures as predictors of first-fixation duration in a more natural text-reading situation. Participants read excerpts from contemporary newspaper articles totaling approximately 2,300 words while their eye movements were recorded.

Method

Stimuli comprised excerpts from 10 British broadsheet newspaper articles covering a broad range of topics. Stimulus presentation and eye movement recording were the same as in Experiment 1, except that each excerpt was formatted into one to four display pages, each containing up to 10 double-spaced lines of text. Each excerpt was followed by an untimed yes/no comprehension question.

Twenty participants from the same population tested in Experiment 1 were paid £5 for participating.

Results and Discussion

Data for the first and last words on a line, words preceded or followed by punctuation marks, the first fixation made on a line, fixations longer than 700 ms, and fixations that did not occur during first-pass reading were excluded from analysis.

Because factors such as word length (e.g., Just & Carpenter, 1980) and launch distance (Vitu et al., 2001) also influence initial-fixation duration, we used multiple linear regression techniques as recommended by Lorch and Myers (1990, Method 3) to first remove variance attributable to these factors, and then independently assessed the abilities of word frequency, transitional probability, and the posterior probability measure to explain the remaining variance. When entered separately into regression equations already containing participants, word length, and launch distance, all three variables were significant predictors, $F(1, 19) = 62.60, p < .0001$; $F(1, 19) = 79.68, p < .0001$; and $F(1, 19) = 81.02, p < .0001$, respectively. The best linear fit was achieved with the re-

gression model incorporating the Bayesian posterior probability ($\beta = -.126$, $R^2 = .159$). Moreover, the addition of frequency to the regression model incorporating posterior probability did not significantly improve model fit, $F(1, 19) = 1.18$, $p > .10$, and the addition of transitional probability resulted in a slight improvement only ($R^2 = .160$), $F(1, 19) = 4.61$, $p < .05$. Thus, the Bayesian model's nonlinear combination of prior and transitional probabilities provided the most parsimonious account of the data.[4] The posterior probability measure captures the notion of lexical predictability as the integration of context-independent and context-dependent statistical information. Figure 2 indicates that the processing advantage for words with high posterior probability is apparent across a broad range of word lengths.

Comparable results were obtained for the other fixation-time measures; posterior probability was the best predictor of single-fixation duration, $\beta = -.136$, $F(1, 19) = 82.67$, $p < .001$. Model fit was not improved by the addition of frequency to the equation already containing posterior probability, $F(1, 19) = 2.70$, $p > .10$, but adding transitional probability resulted in a slight improvement, $F(1, 19) = 6.83$, $p < .05$. The posterior probability measure was also the best predictor of gaze duration, $\beta = -.144$, $F(1, 19) = 99.60$, $p < .001$. In contrast to the results of the initial-fixation and single-fixation analyses, the analysis of gaze duration showed that frequency was a significant predictor even with posterior probability already in the equation, $F(1, 19) = 27.91$, $p < .001$, but transitional probability was not, $F(1, 19) < 1$. This pattern of results suggests that word frequency influences later stages of processing that are reflected in gaze durations but not initial- or single-fixation durations, such as whether the word is refixated (e.g., Rayner et al., 1996).

CONCLUSIONS

The efficiency with which readers can extract the meaning from written language is remarkable (Rubin & Turano, 1992). Cognitive processing strategies and solutions that have emerged over the course of evolution are co-opted to deal with a novel, culturally specific task. Given the existence in readers' input of readily available statistical information about word-to-word contingencies, our findings demonstrate that the brain could exploit this information during reading. Our results fit well with work showing that distributional regularities in the linguistic environment can contribute to the development of basic language abilities. We suggest that the remarkable efficiency of reading is due, at least in part, to the on-line formation of predictions about upcoming words. The statistical properties of the linguistic environment offer a viable source for these predictions.

4. Occam's razor is the basis of our claim of parsimony. The nonlinear Bayesian posterior probability measure is as good a predictor of initial-fixation duration as the linear combination of prior and transitional probabilities arrived at by the regression analysis.
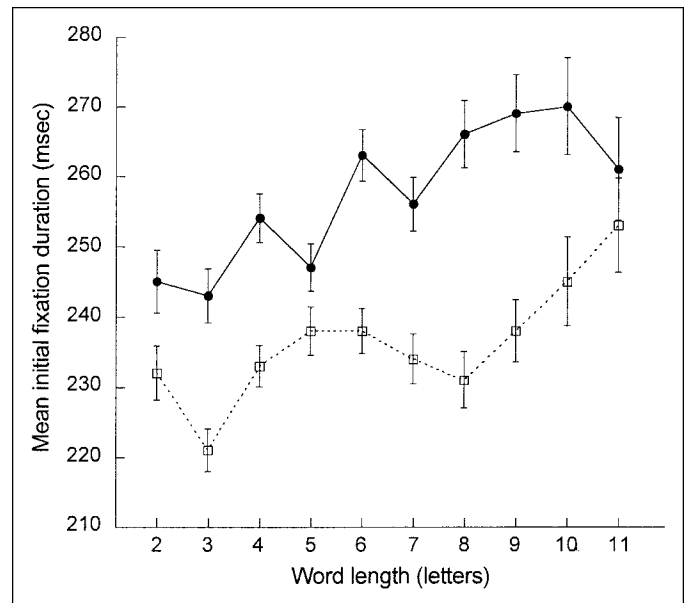

Fig. 2. Mean initial-fixation duration as a function of word length in Experiment 2. The dashed line shows results for the words with high Bayesian posterior probability (defined as the upper quartile of the data); the solid line shows results for the words with low Bayesian posterior probability (the lower quartile). Error bars indicate standard errors.

REFERENCES

Altarriba, J., Kroll, J., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition*, 24, 477–492.

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

Burnage, G., & Dunlop, D. (1992). Encoding the British National Corpus. In J.M. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis, exploitation* (pp. 79–95). Amsterdam: Rodopi.

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.

Inhoff, A.W. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of Verbal Learning and Verbal Behavior*, 23, 612–624.

Inhoff, A.W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40, 431–439.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W.D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: John Benjamin.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–354.

Krauzlis, R., & Adler, S.A. (2001). Effects of directional expectations on motion perception and pursuit eye movements. *Visual Neuroscience*, 18, 365–376.

Kutas, M., & Hillyard, S.A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307, 161–163.

Lorch, R.F., & Myers, J.L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 149–157.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 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, 1188–1200.

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

Rubin, G.S., & Turano, K. (1992). Reading without saccadic eye movements. *Vision Research*, 32, 895–902.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month old infants. *Science*, 274, 1926–1928.

Sereno, S.C., Rayner, K., & Posner, M.I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *NeuroReport*, *9*, 2195–2200.

Taylor, W.L. (1953). "Cloze Procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

Vitu, F., McConkie, G., Kerr, P., & O'Regan, J.K. (2001). Fixation location effects on fix- ation durations during reading: An inverted optimal viewing position effect. *Vision Research*, *41*, 3513–3533.

Zola, D. (1984). Redundancy and word perception during reading. *Perception & Psychophysics*, *36*, 277–284.