Early effects of word surprisal on pupil size during reading

Stefan L. Frank (s.frank@ucl.ac.uk)
Department of Cognitive, Perceptual and Brain Sciences
University College London
26 Bedford Way, London WC1H 0AP, United Kingdom

Robin L. Thompson (robin.thompson@ucl.ac.uk)
Deafness, Cognition and Language Research Centre
Department of Cognitive, Perceptual and Brain Sciences
University College London
49 Gordon Square, London WC1H 0PD, United Kingdom

Abstract

This study investigated the relation between word surprisal and pupil dilation during reading. Participants' eye movements and pupil size were recorded while they read single sentences. Surprisal values for each word in the sentence stimuli were estimated by both a recurrent neural network and a phrase-structure grammar. Higher surprisal corresponded to longer word-reading time, and this effect was stronger when surprisal values were estimated by the neural network. In addition, there was an early, positive effect of surprisal on pupil size, from about 250 ms before word fixation until 100 ms after fixation. This early effect, which was only significant for the network-based surprisal estimates, is suggestive of a preparation-based account of surprisal.

Keywords: Reading; Eye tracking; Pupillometry; Sentence comprehension; Surprisal; Recurrent neural network; Phrase-structure grammar

Introduction

Language comprehension is mostly incremental: When listening to or reading a sentence, each word is immediately integrated with information from the sentence so far (e.g., Just, Carpenter, & Woolley, 1982). It has been argued that the amount of cognitive effort required to process a given word can be quantified by its *surprisal* (Hale, 2001; Levy, 2008), an information-theoretic measure of the extent to which the word's occurrence was unexpected. Formally, if $w_{1...t}$ denotes the sentence's first t words, the surprisal of the following word is: $\text{surprisal}(w_{t+1}) = -\log P(w_{t+1}|w_{1...t})$. These values can be estimated by any language model that assigns probabilities to word sequences.

The relationship between surprisal and cognitive load (i.e., relative difficulty in processing) has indeed been observed in reading studies: Words with higher surprisal values take longer to read, which accounts for several phenomena in sentence comprehension, such as garden-path effects (Brouwer, Fitz, & Hoeks, 2010) and anti-locality effects (Levy, 2008). More generally, reading times at each word in sentences or texts have been shown to correlate with surprisal (e.g., Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Demberg & Keller, 2008; Fernandez Monsalve, Frank, & Vigliocco, 2012; Frank & Bod, 2011; Smith & Levy, 2008).

Here, we investigate an alternative empirical index of cognitive load; one that can be measured continuously and with precise time-resolution: pupil size. By analyzing how and when effects of word surprisal appear in pupillometry data, we are able to use a physiological measure to investigate the fine-grained time course of sentence-comprehension processes.

A large number of studies, using a variety of tasks, have looked at the relationship between cognitive load and pupil dilation (for a recent overview, see Laeng, Sirois, & Gredebäck, 2012). Although these studies differ in how cognitive load is operationalized, increased cognitive load is invariably found to result in larger pupil size. In a non-linguistic context, Preuschoff, 't Hart, and Einhäuser (2011) showed that pupil size (and therefore, presumably, cognitive load) increases when a stimulus is less expected. They had participants perform a simple gambling task and found that experiencing surprise causes pupil dilation: Pupil size correlated not with the gambling outcome itself but with its unexpectedness.

Whether unexpectedness of words in sentences also results in pupil dilation is still an open question. In fact, there has been only very little pupillometry research in psycholinguistics. Engelhardt, Ferreira, and Patsenko (2010) found that a mismatch between syntactic and prosodic structure of auditorily presented sentences results in larger pupil size compared to a condition in which the two structures matched. In another sentence-listening study, Piquado, Isaacowitz, and Wingfield (2010) found a pupil response to both syntactic complexity and sentence length. To the best of our knowledge, there exists only two published studies in which pupillometry is applied during sentence reading: Raisig, Hagendorf, and Van der Meer (2012) presented participants with written descriptions of simple events in everyday activities and found increased pupil dilation when the order of presentation was incongruent with the actual temporal order of the described activities. Just and Carpenter (1993) compared object- and subject-relative clauses and found increased reading times and pupil dilation on the object-relatives, which have long been known to be more difficult to process (Hakes, Evans, & Brannon, 1976). Moreover, the occurrence of a semantically implausible word resulted in increased pupil size compared to a plausible-word condition.

Here, we did not compare particular sentence pairs but, instead, investigate the general relation between word surprisal

and pupil size, looking for effects on each word within a large set of visually-presented sentences. The goals of this study were to explore pupillometry as a methodology for investigating sentence-comprehension processes during reading; to uncover the time-course of surprisal effects; and to assess the suitability of two very different model types for surprisal estimation: recurrent neural networks (RNNs) and phrase-structure grammars (PSGs). We found a very early, positive effect of surprisal on pupil size, which was only significant for the surprisal values generated by the RNN. These findings suggest that surprisal effects are caused by a process of word prediction rather than word integration.

Method

Eye tracking and pupillometry

Materials The self-paced reading study by Fernandez Monsalve et al. (2012) and Frank (2012) used 361 sentence stimuli, semi-randomly selected from three novels published on www.free-online-novels.com. Two hundred and five of these sentences (comprising 1931 word tokens) could fit on a single line of the display and were therefore used in the current eye-tracking experiment. Of those 205 sentences, 110 had a corresponding yes/no comprehension question to ensure that subjects were reading for meaning.

Participants Seventeen monolingual, native English speakers were recruited from the University College London subject pool. One participant was excluded due to technical issues, leaving 16 participants (11 women, mean age 27.6) with analyzable data.

Procedure Subjects were seated 50 cm from the monitor with their chin on a chin rest. Both eyes were tracked using a head-mounted eye-tracker (SR Research, EyeLink II). Individual sentences were presented in 18-point Courier font, left-aligned on the display. Each sentence was preceded by a left-aligned fixation cross that was presented for 800 ms. Gaze direction and pupil area were sampled at a rate of 500 Hz.

After initial calibration (nine fixation points) and five practice trials, subjects were invited to ask clarification questions and the experiment began. Another calibration check was performed after the practice items and then again after every 35 trials (the final set had only 30 trials), at which time subjects took a self-paced break (total 205 trials, six sets). Additionally, drift correction on a single centrally located fixation point was performed at the start of each trial. Responses were recorded using a mouse (center button to continue after finishing a sentence; right and left buttons to respond 'yes' or 'no', respectively, to comprehension questions). The entire experiment (with instructions and calibration) took approximately 50 minutes to complete. The order of trial presentation was randomized throughout.

Surprisal estimation

For each word in the experimental sentences, surprisal values were generated by the same set of probabilistic language models as used by Fernandez Monsalve et al. (2012). All models were trained on 702,412 sentences (comprising 7.6 million word tokens; 7,754 word types) from the written-text part of the British National Corpus.

Recurrent neural network Although RNNs are often used for learning the statistics of language, they are nearly always applied to artificial toy languages. Training such models on a large, English-language corpus, as we do here, requires something more advanced than the standard Simple Recurrent Network (SRN; Elman, 1990). The solution was to first encode each word as a distributed vector and train the network on sequences of those word representations. More precisely, network training was divided into three distinct stages (see also Fernandez Monsalve et al., 2012; Frank, 2012):

1. A co-occurrence matrix $\mathbf{P} = (p_{ij})$ was constructed, where each p_{ij} is the (smoothed) probability that word types i and j occur adjacently in the training data. These values were then transformed into $q_{ij} = \log p_{ij} - \log(\sum_k p_{ik} \sum_k p_{kj})$. The 400 columns of $\mathbf{Q}$ with highest variance were selected, and formed the 400-dimensional vectors for each of the 7,754 word types. This representational space captures the paradigmatic relations between words (e.g., words of the same syntactic category tend to receive similar representations), which boosts generalization to untrained input.

2. The 702,412 training sentences, in the form of word-vector sequences, were given as input to an SRN that learned to predict the vector representation of the upcoming word w_{t+1} after each sentence-so-far $w_{1\ldots t}$. The SRN used standard backpropagation and received the complete training corpus five times.

3. A two-layer feedforward network with 200 hidden units learned to 'decode' the SRN's output vectors into localist representations, that is, into 7,754-dimensional vectors where each element corresponds to a word type. It received the training data two times and, like the SRN, used standard backpropagation for connection-weight update. Its output units have softmax activation functions, so each output vector forms a probability distribution over word types.

The complete model, combining these three stages, generates estimates of the probabilities $P(w_{t+1}|w_{1\ldots t})$ for all word types, from which the surprisal of the actual next word follows directly. These surprisal values were obtained at ten intervals during training of the decoder network, resulting in ten sets of surprisal estimates (by an increasingly well-trained model) each of which was analyzed independently.

Phrase-structure grammar Grammars are usually not induced from 'flat' word sequences but require complete syntactic tree structures as training material. It was therefore necessary to first obtain such structures by parsing the training sentences. This was done by the Stanford Parser (version 1.6.7; Klein & Manning, 2003). The resulting collection of tree structures served as the PSG training corpus.

In a standard probabilistic context-free grammar, the probability of a production rule is conditioned on the rule's left-hand side. For example, the rule 'NP $\rightarrow$ Det N' would be associated with the probability that a phrase consists of a determiner (Det) followed by a noun (N), given that it is a noun phrase (NP). A grammar's structural sensitivity can be increased by also conditioning on other parts of the tree structure, for example, by estimating the probability of 'Det N' given that the current phrase is an NP that belongs to a verb phrase. In this manner, many different grammars, with different structural sensitivities, can be induced from the same set of training data. Here, we applied Roark's (2001) grammar-induction algorithm to obtain eight different grammars (see also Fernandez Monsalve et al., 2012; Frank & Bod, 2011). Next, an incremental parser (Roark, 2001) processed the experimental sentences. At each word, it computed the probabilities of possible syntactic structures[1] (under each of the eight grammars) given the sentence-so-far $w_{1\ldots t}$. The sum of those probabilities equals $P(w_{1\ldots t})$, and surprisal values follow because $-\log P(w_{t+1}|w_{1\ldots t}) = \log P(w_{1\ldots t}) - \log P(w_{1\ldots t+1})$. That is, for each word we obtain eight grammar-based surprisal values, in addition to the ten RNN-based surprisals discussed above.

Results

All participants displayed adequate comprehension by answering at least 80% of the comprehension questions correctly. We excluded from consideration the first and last word of each sentence, clitics, words attached to a comma, the first fixated word, and non-fixated words. Further, data corresponding to fixations outside the sentence presentation region, as well as regressions (i.e., fixations to words earlier in a sentence after a fixation on a later word) were discarded.

Word-reading time

Analysis As a measure of word-reading time, we took total fixation time on a word before fixation on any other word (i.e., the first-pass reading time, or gaze duration; av. 231 ms, s.d. 116 ms). A mixed-effects regression model was fitted to this dependent variable (14,304 data points), using as predictor variables: sentence presentation order (both linear and quadratic factors), word position in sentence (linear and quadratic), word length, log of word frequency, and log of forward transitional probability (i.e., the word's probability given the previous word). Also, all significant two-way interactions were included,[2] as were all significant random slopes of main effects.[3]

The effect size of surprisal is defined as the decrease in regression model deviance when surprisal is included as an additional predictor. This quantity is the χ^2-statistic of a log-likelihood test for the significance of surprisal. Effect size can be contrasted with 'linguistic accuracy': the extent to which the model has learned the statistical patterns of the language. Linguistic accuracy is quantified as the average of $\log P(w_{t+1}|w_{1\ldots t})$ estimated over the experimental sentences, weighted by the number of times w_{t+1} occurs in the analysis.

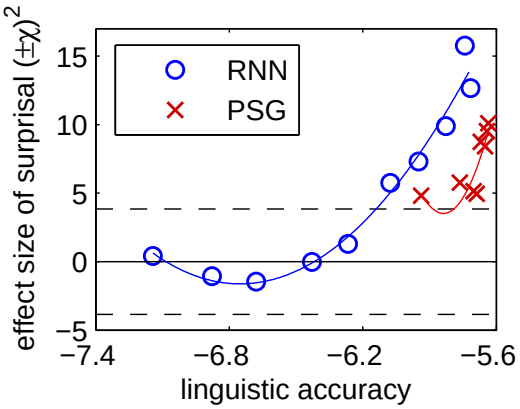

Figure 1: Effect of surprisal (as estimated by either RNN of PSG) on gaze durations as a function of linguistic accuracy (average $\log P(w_{t+1}|w_{1\ldots t})$). Plotted are the estimated χ^2-statistics (where negative values denote effects in the negative direction) and best fitting second-degree polynomials. The dashed lines at $\chi^2 = \pm 3.84$ denote the level beyond which $p < .05$.

Surprisal effect Figure 1 plots the size and direction of the surprisal effect as a function of linguistic accuracy. As expected, all the statistically significant effects are in the positive direction: More surprising words take longer to read. Moreover, models that capture the statistics of the language more accurately also account for more variance in reading time.

Surprisal as estimated by the RNN model (after sufficient training) shows stronger effects than does PSG-based surprisal. We compared the RNN and PSG that showed the strongest effects by testing whether one set of surprisal estimates had an effect over and above the other. The RNN's surprisals did have an additional effect over the PSG's ($\chi^2 = 7.6; p < .01$) but the reverse was not the case ($\chi^2 = 1.93; p > .15$). That is, the grammar does not yield surprisal values that explain any unique variance in reading times.

Pupil size

Analysis As the eyes move across the screen, the angle between the eye gaze and camera changes, affecting the observed pupil size. This was corrected for by fitting a second-degree polynomial to the measured pupil sizes during saccades as a function of the horizontal gaze direction. Correction was performed for each presentation block (i.e., between recalibrations), participant, and individual eye (left or right).

[1] The least probable structures were ignored to make this computation feasible.

[2] These were determined by first including all two-way interactions and then removing the least significant ones until all $|t| > 2$.

[3] These were a by-item slope of sentence order and by-subject slopes of all factors except forward probability and quadratic sentence order.

The fitted values then served as a baseline of pupil size at each horizontal location on the display. If both eyes were successfully tracked, pupil size was averaged over the two. For each subject and sentence separately, pupil sizes were then rescaled to a percentage of the average over the sentence.

The effect of word surprisal on pupil size was analyzed at every 2 ms sample (i.e., the sampling rate of the eye-tracker), from 500 ms before the first fixation on a word, up to 1000 ms after that fixation. If any pupil size during that 1500 ms time window was below 70% or above 130%, the data for those 1500 ms were discarded.

When we analyzed reading times, a baseline regression model was fitted to the gaze durations. In the case of pupil sizes, however, it is not possible to fit just one baseline model because the values of the dependent variable differ across samples. Alternatively, a different model could be fitted to each sample but that would make it impossible to track the surprisal effect over time. Therefore, the same, simplified baseline model is used for all samples. It contained the main effects from the reading-time analysis, except that the factor 'word position' was replaced by the fixated letter's position in the sentence (both linear and quadratic factors). Letter position allows us to take into account differences in luminosity across the display, which can affect pupil dilation. In addition, because samples of pupil dilation are taken up to 500 ms before fixation on the current word, the length, log frequency and log forward probability of the *previous* word are also included. As before, the effect size of surprisal was defined as the decrease in regression model deviance due to surprisal. Surprisal estimates were taken from the RNN and PSG model that explain the most variance in gaze duration.

Surprisal effect Figure 2 shows how strongly a word's surprisal affects pupil size, time-locked to the moment of first fixation on that word. There is a positive relation between surprisal and pupil size, which arises very early, even before fixation (i.e., parafoveally).

Considering that the effect of a word's surprisal arises before fixation on that word, it makes sense to discard cases in which the previous word was not fixated. Specifically, it is unlikely that enough information about a word can be obtained if it is still more than one word ahead. Indeed, as shown in Figure 3, the effect of surprisal remains as strong even when we only take into account cases in which there is a fixation on the previous word (in spite of a 30.3% reduction in the amount of data).

Entropy effect Alternatively, the early effect of surprisal could be due to readers' uncertainty about the upcoming word.[4] If uncertainty about w_{t+1} correlates positively with its surprisal, and being in a state of increased uncertainty causes the pupils to dilate, then the apparent effect of surprisal may actually be an effect of uncertainty. Such an effect can appear during processing of w_t, without any information about the upcoming word w_{t+1}.

We investigated this possibility by estimating how much uncertainty about w_{t+1} a reader may experience after processing $w_{1...t}$. In information theory, uncertainty about the value of a random variable is quantified by its *entropy*. In the context of incremental sentence comprehension, the uncertainty about w_{t+1} is defined as:

$$H(w_{t+1}) = -\sum_{w_{t+1}} P(w_{t+1}|w_{1...t}) \log P(w_{t+1}|w_{1...t}).$$

The entropy $H(w_{t+1})$ is based on the probability distribution $P(w_{t+1}|w_{1...t})$, which is exactly the output of the RNN model. Note that, unlike the word's surprisal, the entropy over w_{t+1} does not require knowledge of the actual upcoming word w_{t+1}. Crucially, $H(w_{t+1})$ equals the expected value of surprisal(w_{t+1}) so the two values correlate positively ($r = .38$ in our data set). A positive effect on pupil dilation of uncertainty about w_{t+1} could therfore be misinterpreted as an effect of the surprisal of w_{t+1}. However, as Figure 4 shows, the relation between entropy (as estimated by the RNN) and pupil size is (if anything) negative. Consequently, the effect of surprisal in Figure 3 is not an entropy effect in disguise.

Discussion

Our reading-time results corroborate earlier findings: More surprising words take longer to read; this effect grows stronger as surprisal values are estimated by linguistically more accurate models; and RNN-based surprisals account for more variance than do grammar-based estimates. Like Frank and Bod (2011), we found no additional effect of the grammar-based surprisals. However, applying the same surprisal estimates to data from a self-paced reading study, Fernandez Monsalve et al. (2012) did find an additional effect of the PSG's surprisals, possibly because their data set was almost ten times larger than our current set.

Importantly, predictions by computational models of language have never before been applied to the analysis of pupillometric data. Hence, the effect of word surprisal on pupil size had not yet been demonstrated. This effect confirms that surprisal is indeed a cognitively relevant measure of processing load, and not merely of processing *time*.

Two explanations have been proposed for the relation between word surprisal and cognitive load: According to Levy (2008), integrating a new word into the interpretation of the sentence so far comes down to updating a probability distribution over all possible sentence interpretations. He proves that the extent of this update, expressed as the Kullback-Leibler divergence from the old distribution to the new, equals the word's surprisal. Alternatively, Smith and Levy (2008) argue that the surprisal effect is due to the reader's processing system being more prepared for more expected words. Under that account, we may expect surprisal effects to occur sooner than if they result from integration of the new word with the current sentence interpretation. Therefore, the very early, pre-fixation effect we found here seems most compatible with Smith and Levy's preparation account.

[4] We would like to thank an anonymous reviewer for this suggestion.

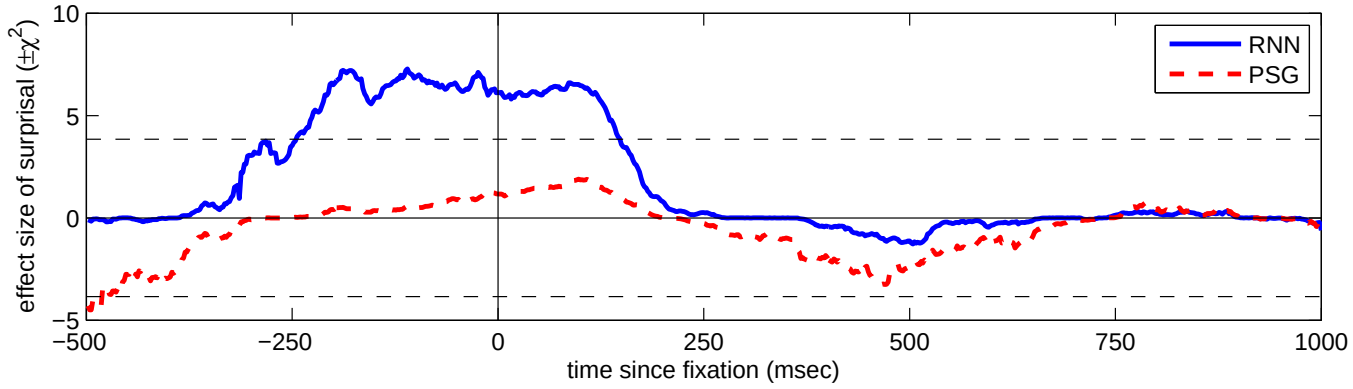

Figure 2: Effect of surprisal on pupil size over time.

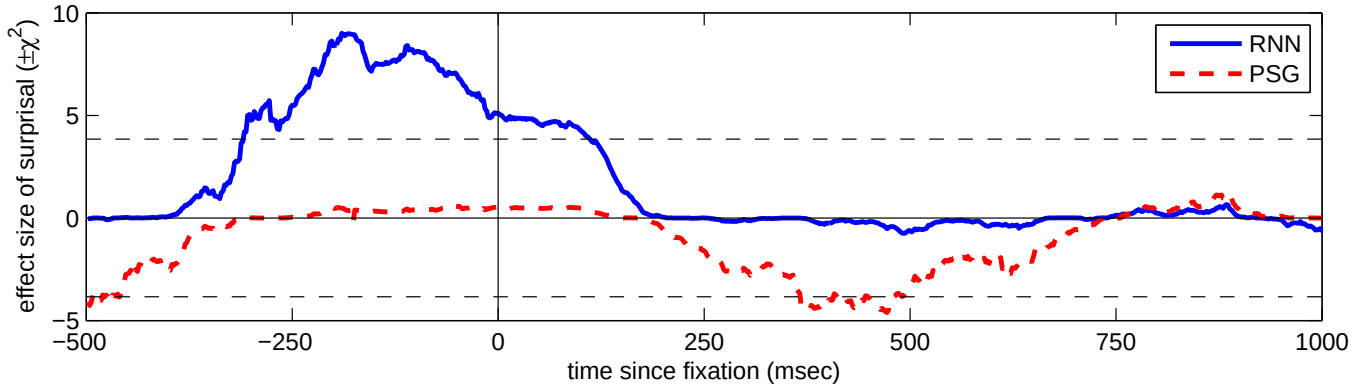

Figure 3: Effect of surprisal on pupil size over time, taking only cases where the previous word was fixated.

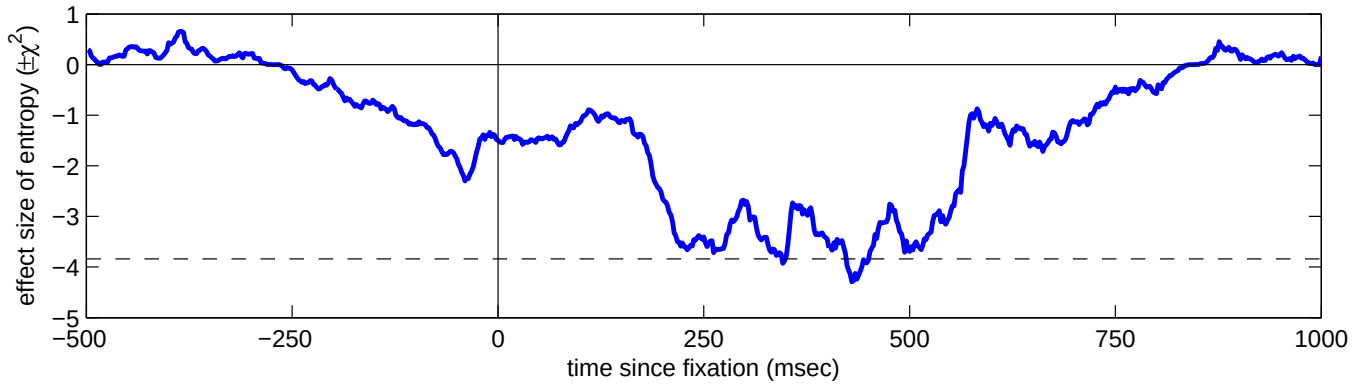

Figure 4: Effect of entropy (uncertainty about the upcoming word) on pupil size over time.

The early occurrence of a surprisal effect may also explain why only the RNN predicted pupil size. Presumably, RNNs simulate early, predictive processing whereas applying a PSG (i.e., parsing) generates syntactic structure and therefore models later 'integrative' processing. Hence, an early effect on pupil size that does not depend on integrative processing would only be predicted by RNNs and not by PSGs.

Conclusion

A word's surprisal has a very early effect on pupil size during reading: At about 250 ms *before* the word is fixated, its surprisal is a significant predictor of the reader's pupil size. This suggests that surprisal effects are due to preparation (Smith & Levy, 2008) rather than integration (Levy, 2008). Moreover, it may explain why surprisal estimates by RNNs have a stronger effect than those from PSGs. Perhaps more importantly, however, we have established that pupillometry is a viable paradigm for studying the fine-grained time course of reading processes.

Acknowledgments

The research presented here was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant number 253803, and by a grant from the Economic and Social Resesarch Council of Great Britain (RES-620-28-6001) awarded to the Deafness Cognition and Language Research Centre. We are grateful to Naima Ansari for her assistance with data collection.

References

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

Brouwer, H., Fitz, H., & Hoeks, J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72–80). Uppsala, Sweden: Association for Computational Linguistics.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, *63*, 639–645.

Fernandez Monsalve, I., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France: Association for Computational Linguistics.

Frank, S. L. (2012). Uncertainty reduction as a measure of cognitive processing load in sentence comprehension. *Manuscript submitted for publication.*

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829–834.

Hakes, D. T., Evans, J. S., & Brannon, L. L. (1976). Understanding sentences with relative clauses. *Memory & Cognition*, *4*, 283–290.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*, 310–339.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238.

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics* (pp. 423–430).

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*, 18–27.

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

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*, 560–569.

Preuschoff, K., 't Hart, B. M., & Einhäuser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, *5*.

Raisig, S., Hagendorf, H., & Van der Meer, E. (2012). The role of temporal properties on the detection of temporal violations: insights from pupillometry. *Cognitive Processing*, *13*, 83–91.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*, 249–276.

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