The relationship between pupil dilation and neural surprise in natural language comprehension

Gehmacher, Q.[1,2], Schubert, J.[3], Kaltenmaier, A.[1,2], Weisz, N.[3,4], Press, C.[1,2]

1 Department of Experimental Psychology, University College London, WC1H 0AP, United Kingdom.

2 Functional Imaging Laboratory, UCL Queen Square Institute of Neurology, University College London, WC1N 3AR, United Kingdom.

3 Paris-Lodron-University of Salzburg, Department of Psychology, Centre for Cognitive Neuroscience, Salzburg, Austria.

4 Neuroscience Institute, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria.

Abstract

Predictive processing theories propose that language comprehension involves generating and updating context-based expectations. We tested whether such predictions are reflected not only in neural activity but also in pupil-linked responses. Using GPT-2, we derived contextual predictions and analysed MEG and pupil data recorded during audiobook listening. Replicating prior work (Heilbron et al., 2022), we find that MEG responses are modulated by both lexical surprise and semantic prediction error. Extending this, we show that pupil dilation selectively tracks semantic prediction error, suggesting sensitivity to meaning-level violations. We assess the mapping function from surprise to these pupil and MEG measures, focusing on linear vs non-linear response profiles and discuss their relation with respect to current predictive processing theories.

Keywords: Language Processing; Locus Coeruleus; MEG; Predictive Processing; Pupil Dilation; Semantic Prediction Error;

Introduction

Theories of predictive processing propose that the brain continuously generates expectations based on context to guide perception and cognition (Friston, 2010; Clark, 2013). In language, growing evidence suggests that such predictions occur at multiple representational levels and shape both behaviour and brain responses.

Recent advances in language modelling have enabled precise estimation of contextual predictions. Using deep neural networks like GPT-2, Heilbron et al. (2022) demonstrated that during naturalistic language comprehension, MEG signals reflect continuous prediction at phonological, lexical, and semantic levels. These findings offer strong support for hierarchical predictive processing accounts.

Here, we analysed MEG alongside concurrent eye-tracking data to ask about the relationship between pupillary responses and these MEG signatures. We replicate the MEG effects reported by Heilbron et al., confirming that both lexical surprise and semantic prediction error modulate brain activity. We then asked whether whether and how prediction error is also reflected in pupil dilation, a peripheral index of locus coeruleus–noradrenergic (LC-NA) activity and arousal (Joshi et al., 2016), and the relation between the MEG and pupillary indices. We examine how surprise may be encoded physiologically by comparing alternative response functions, including non-linear mappings, to assess whether arousal responses scale proportionally with prediction error or reflect context-sensitive gating shaped by relevance or internal model precision - ideas central to theoretical accounts such as the opposing process theory (Press et al., 2020).

Methods

Participants and data We analysed data from 29 participants (12 female, mean age = 25.7 years; Schubert et al., 2024). One was excluded for excessive blinking, yielding a final sample of 28. All had normal hearing/vision and gave informed consent. The study was approved by the University of Salzburg ethics committee.

Stimuli and task Participants listened to four audiobook stories presented in blocks. Stimuli were delivered binaurally at 40 dB above individual hearing threshold. During the task, participants were instructed to keep their gaze directed towards a central cross.

Lexical predictions To estimate contextual predictions, we used the German version of GPT-2, a pretrained transformer-based language model. Raw story transcripts were tokenized and passed through the model using a windowed approach to handle sequences exceeding the 512-token context limit. For each word, we extracted the model's conditional probability and computed lexical surprise as the negative log-probability of each word given its preceding context. We computed semantic prediction error as the cosine distance between the model's expected semantic vector - estimated as a weighted average of predicted semantic embeddings - and the embedding of the observed word.

Control variables We included both acoustic and lexical-semantic control regressors in our analysis. Broadband Envelope: Computed from the gammatone spectrogram (256 bands; 20–5000 Hz) using Eelbrain and summed across all frequency channels. Acoustic Onsets: Derived using an auditory edge detection algorithm applied to the spectrogram,

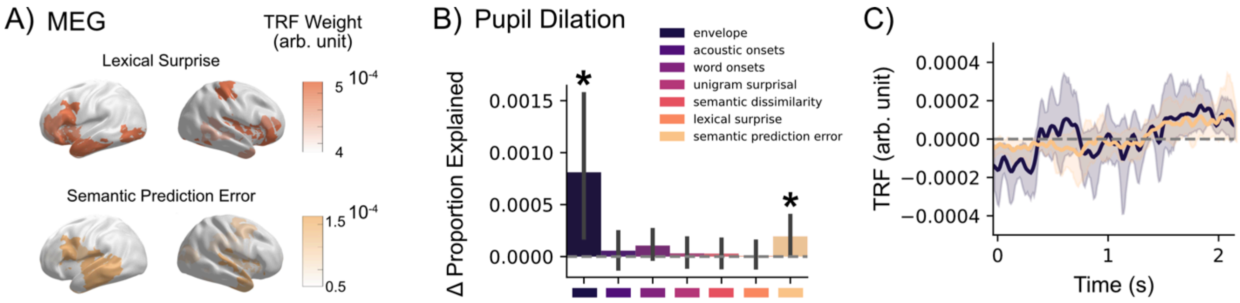

Figure 1: **A)** MEG responses were modulated by both lexical surprise and semantic prediction error, replicating prior findings. **B)** In contrast, pupil dilation selectively tracked semantic prediction error and the broadband envelope, suggesting a dissociation between cortical and LC-NA-linked systems in processing linguistic predictions. **C)** The temporal profile of pupil responses to semantic prediction error peaked around 1.5 - 2 seconds, indicating a delayed, potentially post-perceptual integration process.

capturing sharp temporal changes in the acoustic signal. Lexical Frequency (Unigram Surprisal): Based on word frequency from GloVe embeddings. Semantic Distance: A proxy for integration difficulty, indexing the semantic similarity between each incoming word and its preceding context.

Temporal Response Function analysis To model the relationship between predictors and time-resolved neural and pupillary responses, we used multivariate Temporal Response Function (TRF) analysis via boosting-based deconvolution (David et al., 2007), implemented in the Eelbrain Python package (Brodbeck et al., 2023). Predictors included semantic prediction error, the control regressors listed above, and word onsets. TRFs were computed across all channels and time points.

Results

We first examined the neural correlates of linguistic predictions using MEG data. Replicating findings from Heilbron et al. (2022), we observed robust effects of both lexical surprise and semantic prediction error on brain responses during naturalistic language comprehension. These effects were primarily localized to bilateral auditory regions, consistent with a hierarchical predictive processing architecture (see Figure 1A).

We then investigated whether similar effects were reflected in pupil dilation. Here, we observed a distinct pattern. Pupil responses were significantly modulated by semantic prediction error, but not by lexical surprise or other linguistic control variables, including unigram surprisal and semantic dissimilarity. In addition, we found a significant effect of the broadband envelope. No reliable effects were observed for acoustic onsets or word onsets (see Figure 1B).

Taken together, these results suggest a dissociation between cortical and LC-NA-linked systems: while MEG signals reflect both lexical and semantic predictions, pupil-linked activity appears selectively tuned to semantic-level violations and global acoustic dynamics.

Discussion

Our results reveal a functional dissociation: while MEG responses reflect both lexical and semantic predictions, pupil dilation selectively tracks semantic prediction error. This supports the idea that LC-NA activity is gated by relevance or precision (Bouret & Sara, 2005; Feldman & Friston, 2010; Press et al., 2020), rather than being uniformly driven by surprise.

Rather than signalling generic unpredictability, the LC-NA system may prioritize high-level deviations that challenge stable internal models. Similar distinctions between low- and high-level prediction effects have been noted in attention, where goal relevance outweighs sensory novelty (Corbetta & Shulman, 2002).

Ongoing work explores whether non-linear transformations of semantic prediction error better capture pupil dynamics, shedding light on how prediction is represented across cortical and neuromodulatory systems.

Acknowledgements

This work was supported by a Leverhulme Trust project grant (RPG-2022-358) and European Research Council (ERC) consolidator grant (101001592) under the European Union’s Horizon 2020 research and innovation programme, both awarded to CP.

References

Our Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, 28(11), 574–582. https://doi.org/10.1016/j.tins.2005.09.002

Brodbeck, C., Das, P., Gillis, M., Kulasingham, J. P., Bhattasali, S., Gaston, P., ... & Simon, J. Z. (2023). Eelbrain, a Python toolkit for time-continuous analysis with temporal response functions. *eLife*, 12, e85012. https://doi.org/10.7554/eLife.85012

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201–215. https://doi.org/10.1038/nrn755

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in Neural Systems,* 18(3), 191–212. https://doi.org/10.1080/09548980701609235

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. Frontiers in Human Neuroscience, 4, 215. https://doi.org/10.3389/fnhum.2010.00215

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. https://doi.org/10.1038/nrn2787

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, 89(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028

Press, C., Kok, P., & Yon, D. (2020). The perceptual prediction paradox. *Trends in Cognitive Sciences*, 24(1), 13–24. https://doi.org/10.1016/j.tics.2019.11.003

Schubert, J., Gehmacher, Q., Schmidt, F., Hartmann, T., & Weisz, N. (2024). Prediction tendency, eye movements, and attention in a unified framework of neural speech tracking. *eLife*, 13, RP101262. https://doi.org/10.7554/eLife.101262.1