Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models

Micha Heilbron, **Benedikt Ehinger**, **Peter Hagoort** & **Floris P. de Lange**
{m.heilbron, b.ehinger, floris.delange, peter.hagoort}@donders.ru.nl

Donders Centre for Cognitive Neuroimaging, Radboud University Nijmegen, The Netherlands

Abstract

Prediction in language has traditionally been studied using simple designs in which neural responses to expected and unexpected words are compared in a categorical fashion. However, these designs have been contested as being 'prediction encouraging', potentially exaggerating the importance of prediction in language understanding. A few recent studies have begun to address these worries by using model-based approaches to probe the effects of linguistic predictability in naturalistic stimuli (e.g. continuous narrative). However, these studies so far only looked at very local forms of prediction, using models that take no more than the prior two words into account when computing a word's predictability. Here, we extend this approach using a state-of-the-art neural language model that can take roughly 500 times longer linguistic contexts into account. Predictability estimates from the neural network offer a much better fit to EEG data from subjects listening to naturalistic narrative than simpler models, and reveal strong surprise responses akin to the P200 and N400. These results show that predictability effects in language are not a side-effect of simple designs, and demonstrate the practical use of recent advances in AI for the cognitive neuroscience of language.

Keywords: prediction; language; Transformer; GPT-2

Introduction

In a typical conversation, listeners perceive (or produce) about 3 words per second. It is often assumed that prediction offers a powerful way to achieve such rapid processing of often-ambiguous linguistic stimuli. Indeed, the widespread use of language models – models computing the probability of upcoming words given the previous words – in speech recognition systems demonstrates the in-principle effectiveness of prediction in language processing (Jurafsky & Martin, 2014).

Linguistic predictability has been shown to modulate fixation durations and neural response strengths, suggesting that the brain may also use a predictive strategy. This dovetails with more general ideas about predictive processing (Friston, 2005; de Lange, Heilbron, & Kok, 2018; Heilbron & Chait, 2017) and has lead to predictive interpretations of classical phenomena like the N400 (Rabovsky, Hansen, & McClelland, 2018; Kuperberg & Jaeger, 2016). However, most neural studies on prediction in language used hand-crafted stimulus sets containing many highly expected and unexpected sentence endings – often with tightly controlled (predictable) stimulus timing to allow for ERP averaging. These designs have been criticised as 'prediction encouraging' (Huettig & Mani, 2016), potentially distorting the importance of prediction in language.

A few recent studies used techniques from computational linguistics combined with regression-based deconvolution to estimate predictability effects on neural responses to naturalistic, continuous speech. However, these pioneering studies probed very local forms of prediction by quantifying word predictability based on only the first few phonemes (Brodbeck, Hong, & Simon, 2018) or the prior two words (Willems, Frank, Nijhof, Hagoort, & van den Bosch, 2016; Armeni, Willems, van den Bosch, & Schoffelen, 2019). Recently, the field of artificial intelligence has seen major improvements in neural language models that predict the probability of an upcoming word based on a variable-length and (potentially) arbitrarily-long prior context. In particular, self-attentional architectures (Vaswani et al., 2017) like GPT-2 can keep track of contexts of up to a thousand words long, significantly improving the state of the art in long-distance dependency language modelling tasks like LAMBADA and enabling the model to generate coherent texts of hundreds of words (Radford et al., 2019). Critically, these pre-trained models can achieve state-of-the-art results on a wide variety of tasks and corpora without any fine-tuning. This stands in sharp contrast to earlier (ngram or recurrent) language models which were trained on specific tasks or linguistic registers (e.g. fiction vs news). As such, deep self-attentional language models do not just coherently keep track of long-distance dependencies, but also exhibit an unparalleled degree of *flexibility*, making them arguably the closest approximation of a 'universal model of English' so far.

Here we use a state-of-the art pre-trained neural language model (GPT-2 M) to generate word-by-word predictability estimates of a famous work of fiction, and then regress those predictability estimates against publicly-available EEG data of participants listening to a recording of that same work.

Methods

Stimuli, data acquisition and preprocessing

We used publicly available EEG data of 19 native English speakers listening to Hemingway's *The Old Man and the Sea*. Participants listened to 20 runs of 180s long, amounting to the first hour of the book (11,289 words, ~3 words/s). Participants were instructed to maintain fixation and minimise all motor activities but were otherwise not engaged in any task.

The dataset contains raw 128-channel EEG data downsampled to 128 Hz, plus on/offset times of every content word. The raw data was visually inspected to identify bad channels, decomposed using ICA to remove blinks, after which the rejected channels were interpolated using MNE-python. For all
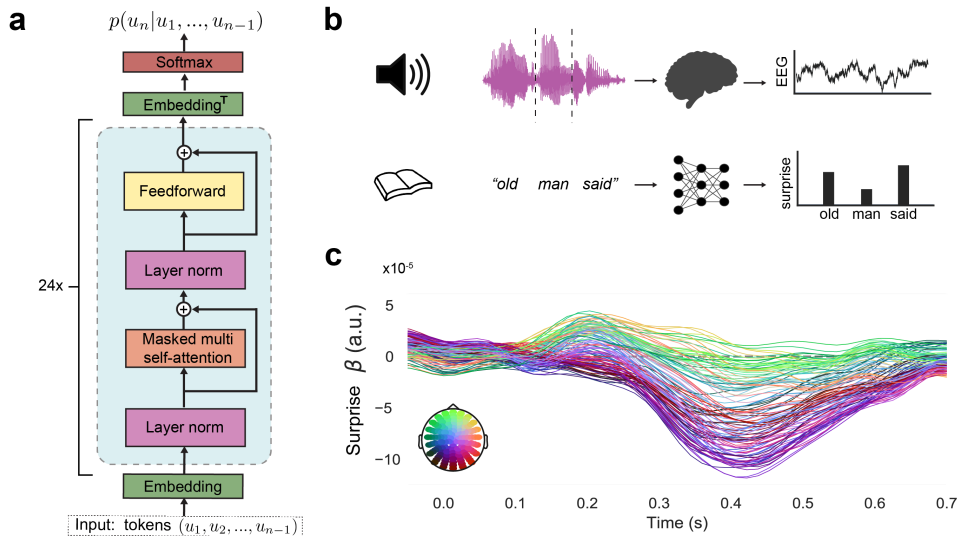
Figure 1: **a)** GPT-2 architecture. For more info on individual operations, see Vaswani et al. (2017). (Note that this panel is a re-rendered version of the original GPT schematic, with subcomponents re-arranged to match the architecture of GPT-2.) **b)** Analysis pipeline overview. **c)** Obtained series of β coefficients (TRF) of lexical surprise (from GPT-2), averaged over participants.

analyses, we focussed on the slow dynamics by filtering the z-scored, cleaned data between 0.5 and 8 Hz using a bidirectional FIR. This was done to keep the analysis close to earlier papers using the same data to study how EEG tracks acoustic and linguistic content of speech; but note that changing the filter parameters does not qualitatively change the results.

For more information on the dataset and prior analyses, see (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018).

Computational models

Word-by-word unpredictability was quantified via lexical surprise – or $-\log\big(p(word|context)\big)$ – estimated by GPT-2 and by a trigram language model. We will describe each in turn.

GPT-2 GPT-2 is a decoder-only variant of the *Transformer* (Vaswani et al., 2017). In the network, input tokens $U = (u_{i-k}, ..., u_{i-1})$ are passed through a token embedding matrix W_e after which a position embedding W_p is added to obtain the first hidden layer: $h_0 = UW_e + W_p$. Activities are then passed through a stack of transformer blocks, consisting of a multi-headed self attention layer, a position-wise feedforward layer, and layer normalisation (Fig 1a). This is repeated n times for each block b, after which (log)probabilities are obtained from a (log)softmax over the transposed token embedding of h_n:

$$h_b = \text{transformer\_block}\,(h_{b-1})\,\forall i \in [1, n] \qquad (1)$$

$$P(u_i|U) = \text{softmax}\left(h_n W_e^\top\right) \qquad (2)$$

We used the largest public version of GPT-2 (345M parameter, released May 9)[1] which has a number of layers (blocks)

of $n = 24$ and a context length of $k = 1024$. Note that k refers to the number of Byte-Pair Encoded *tokens*. A token can be either a word or (for less frequent words) a word-part, or punctuation. How many words actually fit into a context window of length k therefore depends on the text. We ran predictions on a run-by-run basis – each containing about 600 words, implying that in each run the entire preceding context was taken into account to compute a token's probability. For words spanning multiple tokens, word probabilities were simply the joint probability of the tokens obtained via the chain rule. The model was implemented in PyTorch with the Huggingface BERT module[2].

Trigram As a comparison, we implemented an n-gram language model. N-grams also compute $p(w_i|w_{i-k}, ..., w_{i-1})$ but are simpler as they are based on counts. Here we used a trigram ($k = 2$) – which was perhaps the most widely used language model before the recent rise of neural alternatives.[3] To deal with sparsity we used modified Knesner-Ney, the best-performing smoothing technique (Jurafsky & Martin, 2014). The trigram was implemented in NLTK and trained on its Gutenberg corpus, chosen to closely approximate the test set.

Non-predictive controls We included two non-predictive and potentially confounding variables: first, frequency which we quantified as unigram surprise ($-\log p(w)$) which was based on a word's lemma count in the CommonCrawl corpus, obtained via spaCy. Second, following Broderick et al. (2018), we computed the semantic dissimilarity for each content word: $\text{dissim}(w_i) = 1 - \text{corr}\big(\text{GloVe}(w_i), \frac{1}{n}\sum_{i=i}^{n}\text{GloVe}(c_i)\big)$, where

---

[1]For more details on GPT-2, see https://openai.com/blog/better-language-models/ or Radford et al (2019)

[2]see https://github.com/huggingface/pytorch-pretrained-BERT
[3]While $k = 2$ might seem needlessly restrictive, training ngrams beyond $k = 2$ becomes exponentially difficult due to sparsity issues.
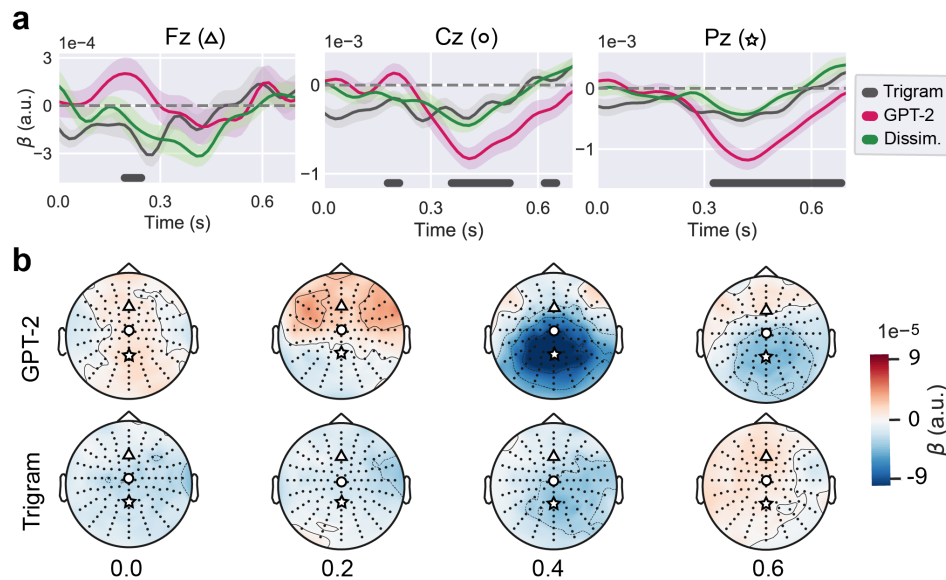
Figure 2: **a)** Grand averaged TRFs for trigram surprise, GTP-2 surprise and semantic dissimilarity for three channels of interest. At each time point, the GPT-2 TRF was compared to both the trigram and semantic dissimilarity TRF with a 2-tailed paired t-test; black bars indicating that both tests were significant at $p < 0.01$, FDR-corrected. Error bars indicate the between-subject SEM. **b)** Topographic maps of grand averaged TRFs for surprise, computed by GPT-2 (top) and the trigram language model (bottom).

$(c_1, ..., c_n)$ are the content words preceding a word in the same or − if w_i is the first content word of the sentence − the previous sentence, and $GloVe(w)$ is the embedding. As shown by Broderick et al. (2018) this variable covaries with an N400-like component. However, it only captures how semantically dissimilar a word is from the preceding words (represented as an 'averaged bag of words'), and not how unexpected a word is in its context, making it an interesting comparison, especially for predictive interpretations of the N400.

Time resolved regression

Variables were regressed against EEG data using time-resolved regression. Briefly, this involves temporally expanding a design matrix such that each predictor column C becomes a series of columns over a range of lags $C_{t_{min}}^{t_{max}} = (C_{t_{min}}, ..., C_{t_{max}})$. For each predictor one thus estimates a series of weights $\beta_{t_{min}}^{t_{max}}$ (Fig 1c) which, under some assumptions, corresponds to the isolated ERP that would have been obtained in an ERP paradigm. In all analyses, word onset was used as time-expanded intercept and other variables as covariates. All regressors were standardised and coefficients were estimated with Ridge regression. Regularisation was set at $\alpha = 1000$ since this lead to the highest R^2 in a leave-one-run-out CV procedure (Fig. 3) Analyses were performed using custom code adapted from MNE's linear_regression module.

Results

We first inspected our main regressor of interest: the surprise values computed by GPT-2, estimated with a regression model that included frequency (unigram surprise) and semantic dissimilarity as nuisance covariates. As can be seen in Figure 1C, the obtained TRF revealed a clear frontal positive response around 200 ms and a central/posterior negative peak at 400 ms after word onset. These peaks indicate that words that were more surprising to the network tended to evoke stronger positive responses at frontal channels at 200 ms and stronger negative potentials at central/posterior channels 400 ms after word onset. Note that while Figure 1C only shows the TRF obtained using one regularisation parameter, we found the same qualitative pattern for any alpha we tested.

We then compared this to an alternative regression model, in which the surprise regressor was based on the trigram model, but that was otherwise identical. Although the TRFs exhibited the same negativity at 400 ms, it was a lot weaker overall, as can be seen from Figure 2B. One anomalous feature is that the TRF is not at 0 at word onset. We suspect this is because 1) we only had onset times for content words, and not for function words typically preceding content words; and 2) for neighbouring words the log-probabilities from the trigram model were correlated ($\rho = 0.24$) but those from GPT-2 were not ($\rho = -0.002$), explaining why only the trigram TRF displays a baseline effect. Further analyses incorporating onset times for all words should correct this issue.

The negative surprise response at 400ms revealed by both the trigram and GPT is similar to the effect of semantic dissimilarity reported by Broderick et al. (2018) using the same dataset. We therefore also looked at the TRF of semantic dissimilarity, for simplicity focussing on the three main chan-
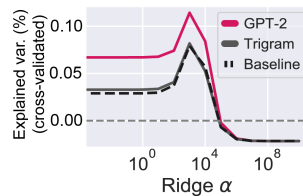
426

Figure 3: Predictive performance of three regression models. We compared a baseline regression model with only unigram surprise and semantic dissimilarity as covariates (dotted line) to two other models that also included surprise values, either obtained from the trigram model (grey) or from GPT-2 (red).

nels of interest analysed by Broderick et al. (2018). At each time-point we compared the GPT-2 TRF to both the trigram and semantic dissimilarity TRF with a 2-tailed paired t-test to find time-points where both tests where significant at $\alpha = 0.01$ (FDR-corrected). As visible in Figure 2b, we observed time-points in all three channels where the GPT-2 TRF was significantly more positive or negative than both other TRFs, confirming that the surprise values from the neural network covary more strongly with EEG responses than the other models.

Finally, to make sure that the difference in coefficients were not related to overfitting or some other estimation problem, we compared the predictive performance of the GPT-2 regression model to the alterntives using a leave-one-run-out cross-validation procedure. As can be seen in Figure 3, this revealed that cross-validated R^2 of the trigram regression model was not significantly higher than that of a baseline model that included only the two nuisance covariates (paired t-test, $t_{19} = -0.25, p = 0.8$); by contrast, R^2 of the GPT-2 regression model was significantly higher than both the trigram regression model (paired t-test, $t_{19} = 5.38, p = 4.1 \times 10^{-4}$) and the baseline model (paired t-test, $t_{19} = 3.10, p = 6.2 \times 10^{-3}$).

Discussion and conclusion

We have shown that word-by-word (un)predictability estimates obtained with a state-of-the-art self-attentional neural language model systematically covary with evoked brain responses to a naturalistic, continuous narrative, measured with EEG. When this relationship was plotted over time, we observed a frontal positive response at 200 ms, and a central negative response at 400 ms, akin to the N400. Unpredictability estimates from the neural network were a much better predictor of EEG responses than those obtained from a trigram that was specifically trained on works of fiction, and than a non-predictive model of semantic incongruence, that simply computed the dissimilarity between a word and its context.

These results bear strong similarities to earlier work demonstrating a relationship between the N400 and semantic expectancy. However, we observed the responses in participants passively listening to naturalistic stimuli, without many highly expected or unexpected sentence endings typically used in the stimulus sets of traditional ERP studies. This suggests that linguistic predictability effects are not just a by-product of simple (prediction encouraging) designs, underscoring the importance of prediction in language processing.

Future analyses will aim at modelling all words, looking at different frequency bands, disentangling different forms of linguistic prediction (e.g. syntactic vs semantic), and trying to replicate these results in different, independent datasets.

References

Armeni, K., Willems, R. M., van den Bosch, A., & Schoffelen, J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*. doi: 10.1016/j.neuroimage.2019.04.083

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current biology: CB*, *28*(24).

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803–809.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. doi: 10.1016/j.tics.2018.06.002

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–836.

Heilbron, M., & Chait, M. (2017). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*. doi: 10.1016/j.neuroscience.2017.07.061

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). Pearson London.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, *31*(1), 32–59.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693. doi: 10.1038/s41562-018-0406-4

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*, 8.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6).