Modelling EEG responses to spoken narration

Rasyan Ahmed[a], Willem Zuidema[a], Tomas O. Lentz[a,b]

[a] *Institute for Logic, Language and Computation, P.O. Box 94242, Amsterdam, 1090GE, Netherlands*
[b] *Tilburg center for Cognition and Communication, P.O. Box 90153, Tilburg, 5000LE, Netherlands*

Keywords: cortical entrainment, semantic processing, computational linguistics, EEG, ERP, contextual embeddings,

1. Summary

When listening to a spoken narrative, listeners construct representations of the story and its characters and events, and integrate the information conveyed by each incoming word with their representation of the preceding context. Decades of research have revealed signatures of this process in brain activity as measured from outside the skull, and have shown that the compatibility of the new information with existing representations and expectations affects these signatures. In particular, Kutas and Hillyard [1] discovered a specific pattern in EEG recordings – later termed the N400 effect – approximately 400 milliseconds after a semantically surprising word (e.g., the word "socks" presented after "coffee with"). Very many studies since that initial discovery have revealed details of the conditions under which the effect can be observed; all of this work relies on the Event-related Potential (ERP) paradigm, involving careful (manual) selection of stimuli, and binary contrasts between a target condition and a control condition.

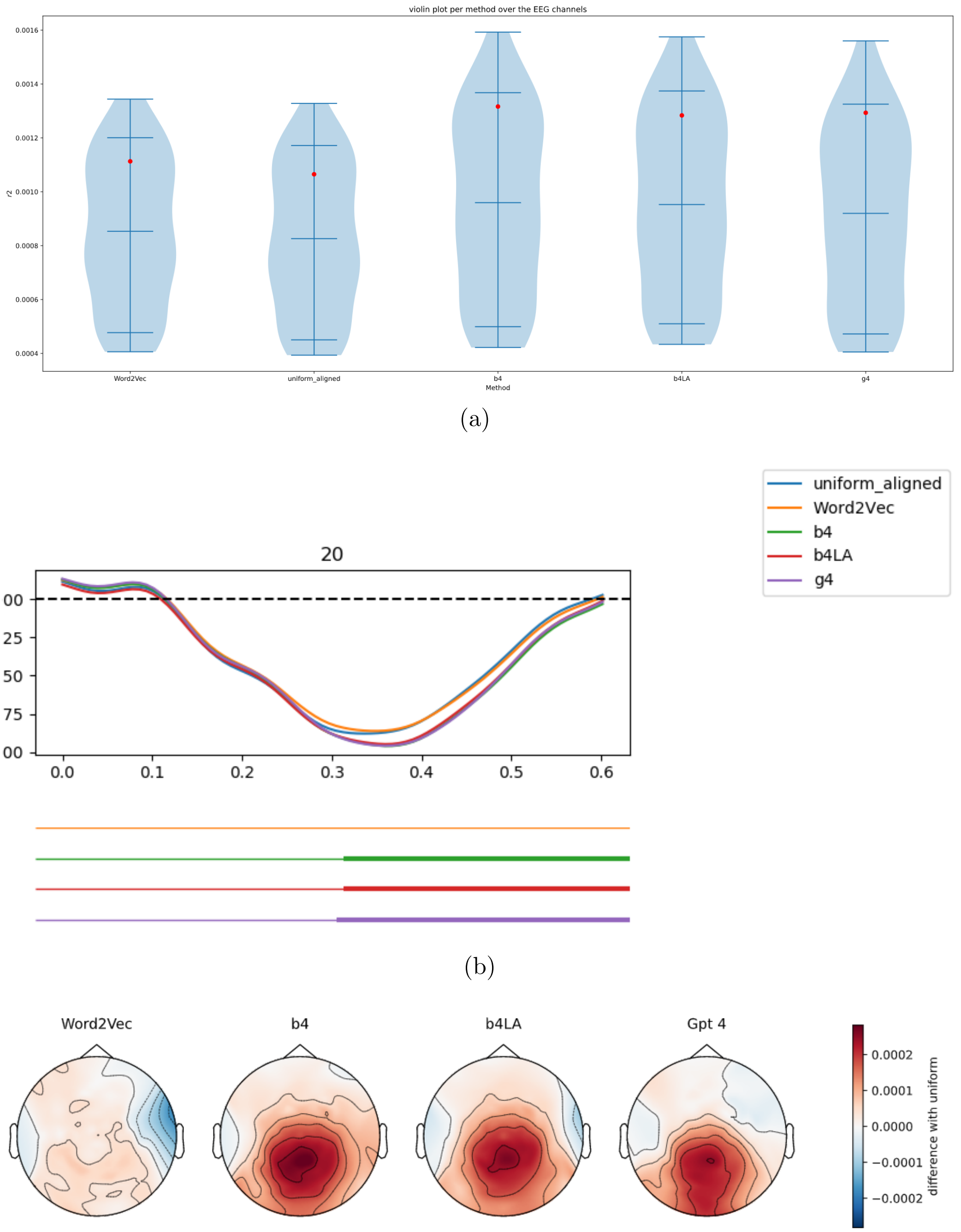
The current study also aims to show an N400 effect, but unlike prior ERP work it uses EEG data gathered in a naturalistic setting (participants listening to an audio book) – without a control condition – and relies on state-of-the-art computational language models from the field of Natural Language Processing to quantify compatability. We build directly on previous work [2], and use the same EEG dataset; crucially, however, we replace their baseline and compatibility measure. We find that the baseline used in this prior work is appropriate for showing an N400 component (a signature of semantic pro-

cessing), but not for the N400 effect (a modulation of the strength of the component). Moreover, we find that our measure of compatability, based on the Transformer language models, explains a much larger fraction of the variance than other measures, and that the predictability of the EEG signal is strongest for the most surprising words and in regions traditionally associated with the N400 effect. This paper thus presents the first evidence that language models in a naturalistic setting can predict both the N400 component and the amplitude of the N400 effect, and their temporal and spatial characteristics.

2. Results and Discussion

We use the dataset of EEG signals recorded from participants listening to an audiobook released by [2]. We extracted word onset times from the stimuli, and postprocessed the EEG recordings using standard procedures. We also reconstructed the compatibility measure from [2], based on the word2vec word vectors for English words [ref]. Additionally, we computed word probabilities of all words in the stimuli according to two state-of-the-art, pretrained language models: BERT and GPT-2. For BERT, we used the BERT-large pretrained model, and considered 2×5=10 varieties: without ($f = 0$) or with ($f = 1$) considering future context in the text, and with a context window within the current sentence only ($h = 0$), or with an additional $h = 1$, 2, 3 or 4 sentences (if $f = 1$, h sentences of future sentences are also included). For GPT2, we used the GPT-large pretrained model with $h = 4$.

We define a baseline model referred to as the UNIFORM-ALIGNED model, that only makes use of word onset times (and assumes a uniform compatability score $c()$ for every word w: $c(w) = 1$). The baseline model can itself be compared to an UNIFORM-UNALIGNED, i.e., an attempt to predict EEG activity with the same amount of spikes, but without word onset information. The UNIFORM-ALIGNED model predicts more of the EEG signal, indicating that word onsets are correlated with brain activity. We compare the UNIFORM-ALIGNED baseline with the WORD2VEC model of [2] ($c(w) = cos(\hat{v}(w'), v(w))$, where $v(w)$ gives the word2vec vector of the focal word, and $\hat{v}(w')$ the average vector of prior words in the sentence). With the modern neural language models, we compute the conditional probability, and use the reciprocal of that probability as compatability score: $c(w) = 1/p(w|h)$, where h is the conditioning context.



(a)

(b)

(c)

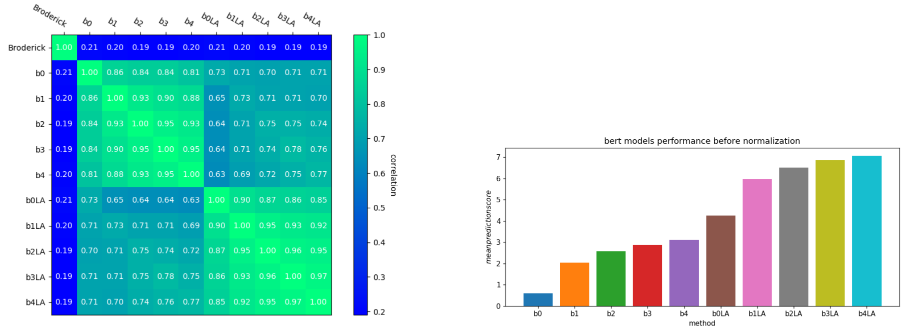
Figure 1: Top: Distribution over channels of explained variance, for each computational model and the UNIFORM-ALIGNED baseline. The explained variance for Pz, where the N400 component is expected, is marked by a red dot. Middle: The Time-Response Function (TRF) for Pz, for a selection of computational models. The lines under the graph show for each model when it is significantly different from the UNIFORM-ALIGNED model; only if the line is thicker there is a significant difference between the TRF values calculated for that time point (calculated on five-fold cross-validation for 19 participants). Bottom: Topographical map of explained variance per electrode location of a selection of computational models, compared to the UNIFORM-ALIGNED baseline model; variance was mainly explained in the posterior regions associated with the N400 component.

Figure 1a shows the main findings. Importantly, we find that, in terms of explained variance in the EEG signal, the WORD2VEC model does not significantly improve over the UNIFORM-ALIGNED baseline (although it does do significantly better than the UNIFORM-UNALIGNED and REVERSED-SPEECH baselines; see star methods for an explanation of why these baselines are less relevant here). This means the findings reported in [2] are driven by knowledge of word onset times, and thus only capture the N400 component. In contrast, the measures based on modern language models all do significantly differ from the baseline. The highest predictability is obtained with models that take at least 4 sentences of prior context into account when computing a word's compatability. The modulation of the brain response based on the individual word's compatability is, in essence, the N400 effect.

Figure 1b shows the TRF curves estimated for the different models for the Pz channel. The shape of the curves are similar to each other and to N400 curves obtained using the ERP paradigm. The figure again shows no significant difference between word2vec and uniform-aligned; the curve for the Bert model is more pronounced and differs significantly around 400 ms. Figure 1c shows, for each model compared to UNIFORM-ALIGNED, the difference in r^2 values for all channels, averaged between 300 and 450 ms after word onset; it shows EEG predictability is centered just below the Pz channel, and again illustrates BERT and GPT2 compatability measures yield results in line with existing work from the ERP paradigm.

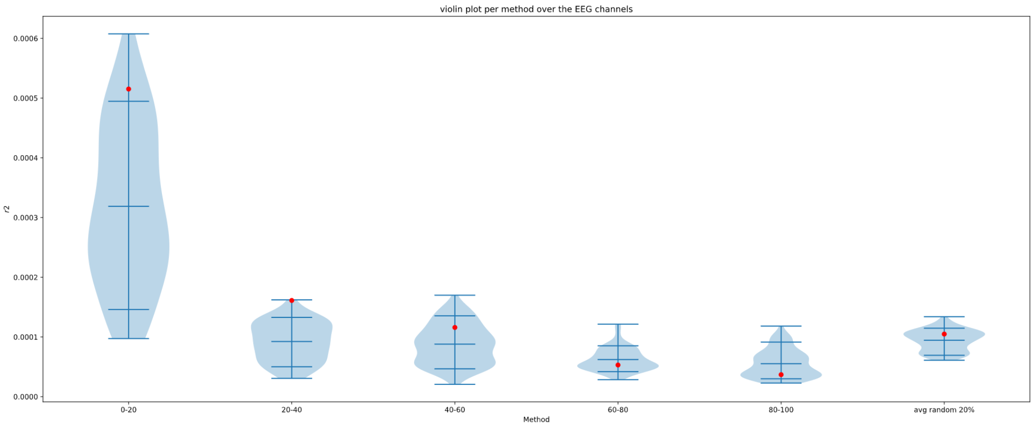
To understand better the differences between the models considered, we compare in Figure 2 the prediction made my these models. Panel (a) shows the correlations between word compatibility scores computed by each of the models. The figure show the word2vec model differs strongly from the other models considered; the second most striking difference is between Bert models that do consider future context (so-called 'masked language modelling') and the cognitively more plausible Bert and GPT models than do not consider future context (so called 'causal language modelling'). Interestingly, the panel (b) shows that models that do have access to future context are much better at predicting the target word. Such improved word predictions did not, however, translate in a better predictability of the EEG response (see Figure 1a).

Finally, in figure 3 we use our setup to study a question that would require considerably more resources to be studied in the ERP paradigm: how does the strength of the N400 effect depend quantitatively on word compatibility? We divide up all the words used in the stimuli into five bins based on the



(a) (b)

Figure 2: Prediction accuracy of different models. Left: Correlations between compatibility score for each word, showing that word2vec only has a weak correlation with the BERT and GPT-2 models. Predictions with and without looking to upcoming context also differ. Right: Comparison of prediction score for actually presented word; the more sentences the Bert model receives, the better it is at predicting the upcoming word. With access to words that come after the word to be predicted, performance increases even more.



(a)

Figure 3: Performance of BERT-4 model on each quintile of the data (quintiles based on compatibility with the context). Singling out the 20% least compatible words allows to partially predict the EEG signal, while the performance for the lower quantiles is comparable to the baseline model. The rightmost violin shows the performance for a random 20% of the data; this amount of data seems to be too low generally for good model fitting. The red dot marks the performance for Pz.

BERT($h = 4$) compatibility scores, and run our TRF estimation algorithm on each of these bins separately. For comparison, we also run the algorithm on a bin containing a randomly selected 20% of the words. As the figure shows, the model only predicts variance when trained and tested on the least compatible (most surprising) words. The relation between compatibility and explained variance appears to be rather nonlinear (tending quickly towards zero explained variance even for moderately compatible words).

3. Discussion

Estimates of word compatibility in a given context as derived from current, transformer-based computational methods such as Bert and GPT2 allow to make the dichotomy of the classical N400 experimental setup continuous. Instead of manually marking words as surprising or unsurprising, surprisal, or its reverse, context compatibility, can be estimated for each word in each context. This richer information allows us to distinguish the N400 effect from the N400 component and yields an initial estimate of when the N400 effect arises.

Crucially, TRF fitting allows to estimate how well an estimate of compatibility correlates with an ERP signal. The Uniform-Aligned model shows that there is a component of the signal that can be predicted with information about the presence of a word. A control condition, in which EEG is recorded while no word is presented to the participant, was not needed to find this component. However, a model that takes into account both the presence of a word and its compatibility with the context can predict even more of the ERP signal. Comparing it to a baseline model that captures the N400 component in general allows to see if the context compatibility measure is informative beyond the N400 component itself. It turns out that the values obtained from word2vec do not explain significantly more variation than the baseline model does. In addition, the calculated TRF shows no significant difference with Uniform-Aligned. As both models perform as well and in a very similar way, the compatibility estimates of word2vec may not reflect the cognitive process and brain activity of human language processing. However, the predictions from Bert and GPT2 do capture more variance than the baseline model and in a different way.

The TRFs for BERT in Figure 1b indicate that the predictions of the ERP signal depending on context compatibility fit at a later time point than the N400 component captured by Uniform-Aligned and word2vec. The later peak is also deeper. In other words, a later dip is predicted for less compatible words, which leads to better predictions than if only the N400 component was predicted. In other words, a modulation of the N400 component can be detected, which amounts to an N400 effect. This effect is, throught the nature of the linear fitting of TRFs, linearly increases with the inverse loglikelihood of a word. Nevertheless, the largest effect is found on the time scale of the 20% most surprising words. The TRF approach is therefore already succesfull on this 20% of the data, even though it generally is not. The effect captured is carried by the most surprising words.

4. STAR ⋆ Methods

The compatibility estimates based on the WORD2VEC model, as used by [2], are based on the difference between the average semantics of preceding words. Estimates based on BERT and GPT-2 depended on predictions of upcoming words and differ substantially from the WORD2VEC predictions. Figure 2 shows that models of similar architecture have similar compatibility measures.

The information contained in the probability BERT assigned to a word given the context is combined with the temporal information contained in the word onset times.

The variation in compatibility with the previously produced words can be used to predict between-word differences in EEG activity, which is illustrated by using the BERT model [ref], a Transformer model trained on 'masked language modelling' on 300 billion words of English text. We consider a number of different choices for two parameters of BERT: (i) amount of prior context provided to the BERT model; (ii) whether or not the remainder of the sentence (future context) is available to the BERT model.

4.1. Model fit

TRFs can be predicted by convoluting a vector of spikes that are time-aligned with words. We replicated TRF modelling of participants listening to an audiobook using a simple derivative of distributional semantics of words [2], time-locked to the onset of the words. The TRFs we fitted indeed negatively correlate with the suprisal measure.

4.1.1. Quantification and statistical analysis

The TRF waveforms were submitted to statistical tests against the hypothesis that they were equal to zero. For each datapoint (a time bin for a participant for an electrode), a TRF was fitted five times. We generalise over participants and fits.

Appendix A. TRF method

The TRF method is visualized in figure A.4.

Appendix B. Highlights

- ERPs, specifically the N400 effect, can be predicted with computational models of contextual embeddings.
- Ecologically valid semantic models predict brain activity better.
- The N400 effect may not be a linear modulation of the N400 component.

References

[1] M. Kutas, S. A. Hillyard, Reading senseless sentences: Brain potentials reflect semantic incongruity, Science 207 (4427) (1980) 203–205.

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

TRF methodology visualised

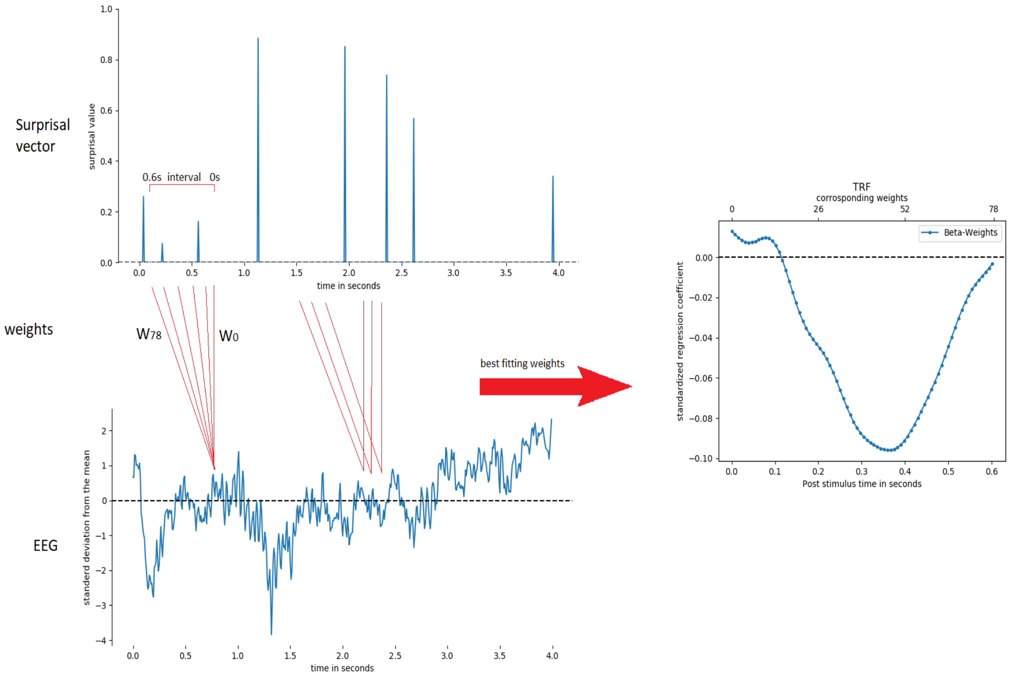


Figure A.4: Shown here is a simplification of the process. Up top are shown a portion of the surprisal vector (SP) and the same portion of the processed EEG. The surprisal vector is a sparse vector where each non zero represents the onset of a word with its height equal to the surprisal value of that word. For each data point in the EEG, the linear regression model tries to predict that point using the 0.6 seconds that came before it on the dissimilarity vector as its input. Each input data point has its own weight that tells it how important that point is for the prediction. The regression model finds the best fitting weights, those that minimize the mean squared error compared to the real EEG. These final weights are then plotted as an TRF.