Modeling EEG responses to narrative speech

Rasyan Ahmed
10784063

Master Thesis
Credits: 36 EC

Master's Artificial Intelligence

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervised by:
dr. Tom Lentz
dr. Jelle Zuidema

8th December 2020

Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Faculty of Science
Science Park 107, 1098 XG Amsterdam
https://www.illc.uva.nl/

Abstract

We expand upon previous work by Broderick et al. which found a negative correlation between how surprising a word is and the EEG data following the onset of that word. This negativity which was only present when participants were semantically processing the spoken text, showed a lot of similarities to the N400 in classical literature. We show by adding additional baselines that what they found is similar to the N400 component but not the N400 effect. Evidence of the N400 effect can be found by moving away from dissimilarity based surprisal calculations to predictability ones through the well known BERT language model. Additionally, we also investigate the model fit of the simplistic TRF model which we suspect is underfitting the complex EEG dataset.

Contents

1 Introduction

Brain imaging techniques such as EEG allow us to glimpse at the electrophysiological processes ongoing during a variety of stimuli, but the large amount of data produced makes it difficult to interpret what is actually happening. Experiments using these kinds of data often employ strict experimental setups. One major finding using such a setup is the N400.

The N400 component, often associated with semantic processing of words, is as indicated by its name, a negative deflection observable in EEG that occurs 400ms after the onset of a word. The N400 effect is an expansion upon the component which links the size of this deflection to how surprising a word is. In most N400 studies, sentences are presented one at a time which are handcrafted such as to maximise how surprising a target word is so it can be compared to a control sentence. But how well do these N400 studies generalise to more realistic situations that occur in everyday life? To study this, we need stronger statistical and machine learning methods that can make sense of the large amount of EEG data for us, so that we can relax the experimental setup.

An attempt to achieve this is the 2018 paper published in Current Biology by Broderick et al. (Broderick, Anderson, Di Liberto, Crosse, & Lalor, 2018). They found evidence of semantic processing in continuous EEG recordings from participants simply listening to an audiobook. Using a linear regression model, a temporal response function can be constructed which shows a negative correlation between the EEG recording after the onset of a word and how surprising that word is in its sentence. In their experiments they showed that this correlation is only present when the participants are able to follow the spoken text. Their results suggest that the correlation captured is therefore likely related to the semantic processing of words, similar to the N400 component. A point of confusion for us is the disconnect between the methodology employed and the experimental setup used to verify it. The use of surprisal calculations suggests a search for the N400 effect, yet the experiments employed focus solely on evidence for semantic processing, which is more related to the N400 component. In what context then do we place the resulting correlation? We note that the authors themselves refrained from calling the correlation itself either the N400 effect or component . In part due to how their results peak slightly before 400 ms post stimulus and due to the theoretical differences between how they calculate surprisal compared to those employed by more classical N400 studies.

In this paper we show that besides the theoretical, there are also more fundamental flaws in their methodology. We remedy these flaws by introducing a new way of calculating how surprising a word is that is theoretically much closer to more classic N400 effect studies[1]. The test suite is expanded with additional baseline in order to see whether such surprisal measures are actually required to recreate the N400 component and whether

[1] A link and explanation of the codebase is provided in Appendix B

they can show evidence of the N400 effect. Lastly we look at the model fit to get an idea of how much signal is captured from the EEG data by the model, as these were missing in the original study.

2 Literature Review

2.1 N400

The N400 was found by Kutas and Hilyard during a study on semantic violations in text (Kutas & Hillyard, 1980). They found that when presenting words one at a time, when the sentence ends with a semantically non congruent word, a large negative spike in the central-parietal scalp sites was detected compared to the reference electrodes placed behind the ears (mastoids). This phenomenon was called the N400 as this negativity seems to be maximal around 400ms post stimulus. In follow up studies the effect was shown to be very robust with respect to the medium the text was presented in, reading, listening, signing and all other kinds of presentation methods were shown to elicit the N400 (Kutas, Neville, & Holcomb, 1987). As the initial study was centred around semantic violations, the N400 was largely linked to these kinds of violations, but later studies showed that even congruent words show moderate N400s (Fischler, Bloom, Childers, Roucos, & Perry Jr, 1983).

It is here that an important distinction needs to be made. The N400 component is the deflection that occurs linked to the processing of words regardless of congruence at roughly 400ms post stimulus. The N400 effect on the other hand is the difference in amplitude of the N400 spike based on congruence. Many factors have since been found that influence the N400 amplitude, which all seem to be related to how expected or unexpected the word is. It was shown that words that appear less frequently in general show larger N400s than more common words (Rugg, 1990). Words that were repeated frequently showed increasingly smaller amplitudes (Rugg, 1985). And words that were primed in verbal fashion through audio or visual stimuli (Holcomb & Neville, 1990) or non verbal fashion such as gestures or pictures (Barrett & Rugg, 1990) have smaller amplitudes compared to when they were not primed beforehand. The strongest correlation found so far (up to r=.9) (Kutas & Hillyard, 1984), is that between the amplitude of the N400 and the Cloze probability of the word in its context. In the Cloze test, sentences are shown with one or more words blanked out and participants are asked to fill in the blanks. The Cloze probability of a word is then defined as the percentage of people that gave that word as the answer to a blank. The previous listed factors are all indirectly incorporated inside the Cloze probability and so the high correlation here is the strongest argument for the link between the N400 and the predictability of a word in its context.

2.2 Temporal response function

The idea of using mathematical functions to model the direct relationship between stimuli and brain response as done here, is not new and commonly called system identification in the field (Marmarelis, 2004). An important subset of the SI methods are Linear and time-invariant. Two assumptions are made in this model, namely that there is a direct linear relationship between input and output and that this relationship is independent of the time. Both of these assumptions are made to simplify the model, and even though they are not applicable to the actual human brain, they can be shown to be reasonable in certain cases (Boynton, Engel, Glover, & Heeger, 1996). We follow an LTI method based on linear regression called the temporal response function (TRF) outlined by the mTRF Matlab toolbox (Crosse, Di Liberto, Bednar, & Lalor, 2016). A recreation of their methodology was made in python with the use of MNE python package (Gramfort et al., 2013).

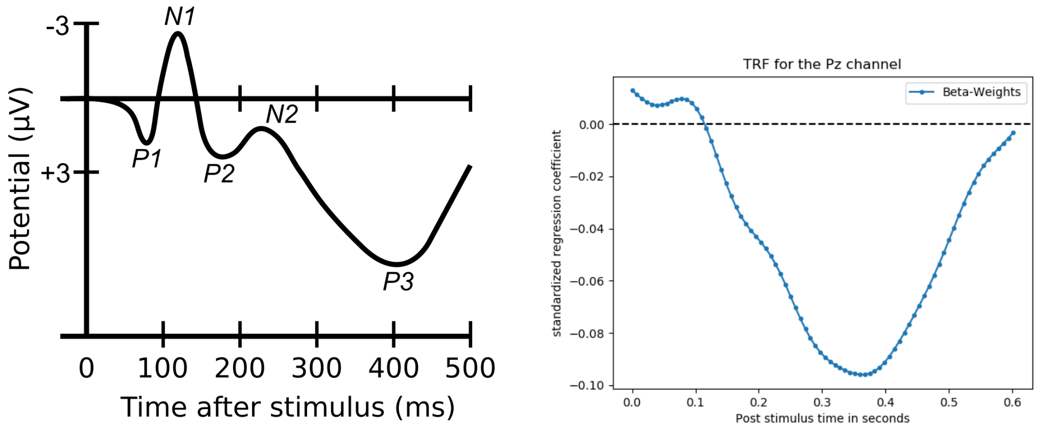


Figure 1: ERP on the left side and TRF on the right side. They are similar in what they show, but made through different methods. The ERP are the raw values where the TRF is trying to reconstruct the best fitting ERP through ML. each individual beta weight on the TRF is plotted as a dot here.

A TRF can be seen as a reconstructed Event related potential (ERP) through the use of for example, regularized linear regression. Figure 1 shows both of them side by side. On the left we see the normal ERP where the event happened at time 0 and we see how the potential develops over time. The same is true for the TRF, yet there is a clear difference. In the ERP one only needs to know when an event happens in order to start measuring from there. In the TRF you have additional information from the event you need to take into account, in our case, how surprising that word was (each content word is its own event). This is reflected in the Y-axis. The TRF shows the predicted standardised regression coefficients instead of the raw electrical potentials measured in the brain. These regression coefficients are the optimal weights found by the regression model and tell you the expected effect of an event on each EEG time point post stimulus.

As such they can be interpreted as the correlation between the surprisal value of the event and the EEG following it. The LTI properties are also encoded within these weights. They show a linear relationship between event and EEG, hence the expected change of an event with half the surprisal value of another event would also be half. The time invariant property is shown in how the weight of each time point (shown as a dot each) is independent compared to all the other time points. For clarity reasons, we will plot these independent time points as only a line from here on out.

2.3 Encoding the meaning of words

In order to evaluate whether a word is surprising or not we need to encode in some mathematical fashion the semantic properties of words. Fortunately, there is a vast field of research by linguists and computer scientists focused exactly on this problem. This field is largely based on the distributional hypothesis outlined in the 50s (Harris, 1954), which notes that words with similar semantic meaning often occur in similar contexts. Using this property, one can then focus on encoding the contexts of a word into a vector with the hope of creating a vector space where similar words are close to each other. These embeddings were often made by counting the number of times other words occur near it and applying a variety of mathematical and linguistic transformations to these co-occurrence counts. As of recently, count based methods have fallen out of flavour as prediction based methods are taking over (Baroni, Dinu, & Kruszewski, 2014). These methods have at their heart a neural network that either predicts a word given its context or the other way around. A well known example is word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), of which we use a pre-trained variant (Baroni et al., 2014).

2.4 From embeddings to language models

Sentences are structured combinations of words in which order matters. How then do you combine the embeddings of the words together to form an encoding of the sentence, a paragraph or document? Averaging word embeddings is not ideal as the average of two vastly different sets of words could be the same. Concatenating the word vectors together and having that as input seems like a much better solution, but here a new problem occurs. Sentences have different lengths and can be very long, making computations complex. Such concatenation based models are called N-gram models and are often used as language models, where the goal is to predict the next word in a sentence (Kneser & Ney, 1995). A predefined number (the N in N-gram) of previous words are given as input which are used to generate a probability distribution for what the next word might be.

To address these problems, a recurrent neural network can be used instead which is sensitive to the temporal axis in text (Rumelhart, Hinton, & Williams, 1986). Such a model is fed word embeddings one by one, and changes its internal representation of the sentence so far. The internal representation can be used at each time step to create a prediction for the next word. While non-static, the context window offered by a RNN is still relatively small as previous words tend to be forgotten quickly in a large corpus. Improvements to the main RNN structure such as in LSTMs (Hochreiter & Schmidhuber, 1997) help in this aspect, but the main solution is to use a different architecture. Transformers do not require words to be fed in sequence Instead, a mechanism is trained that tells the models how important each previous word is in the corpus as references for generating the new word (Vaswani et al., 2017). These attention mechanisms can take into account word relationships much further in the past than previous models.

BERT, Bidirectional Encoder Representations from Transformers, uses as defined by its name a transformer architecture (Devlin, Chang, Lee, & Toutanova, 2018). Instead of training on a next word prediction task it uses masked language modelling where one predicts the original word that was masked in the sentence. The choice for this task is directly inspired by the Cloze task mentioned above. This similarity to the Cloze task and probability's makes BERT uniquely suited as the model of choice for this study.

2.5 Combining NLP models and brain imaging data

A famous example of the use of brain imaging data in tandem with word embeddings is a 2008 paper by Mitchell et al. (Mitchell et al., 2008a). Count based word embeddings were used as inputs to a linear regression model to predict the fMRI activity in each part of the brain when thinking about the word the embedding represents. Using this model, it is then possible to predict with accuracies significantly above chance, which fMRI image belongs to which word embedding. Further research showed that moving to prediction based word embeddings such as word2vec significantly improves the accuracy of the model in the same task (Abnar, Ahmed, Mijnheer, & Zuidema, 2017).

A recent MIT paper builds further upon these results. They show that a variety of NLP models are able to predict with high accuracy fMRI brain activations (Schrimpf et al., 2020). Where prediction based word embeddings produce accuracies below 20%, transformer based models instead are capable of near-perfect accuracy relative to the noise ceiling. The same paper also shows a large correlation between how good a model is at predicting the next word in a sequence and how strong it is at predicting brain activations. This correlation, not present in other NLP tasks, can partly explain the strong performance of language models such as BERT, as they are naturally well suited to this task.

3 Critical replication of Broderick et al.

Broderick et al. applied a methodology that is capable of capturing a negative deflection similar in shape and location to the N400 component in continuous EEG recordings (Broderick et al., 2018). This showed that it is possible to capture evidence of semantic processing through the use of a loose experimental setup where subjects simply had to listen to an audiobook during recordings. The correlation was found through the use of a

TRF created with the dissimilarity of words compared to their context as its only input.

Three experiments were done in total. In the first experiment they compared their findings to a baseline created from measuring the EEG of the subjects while listening to the same audio book but in reverse. They showed that not only is the correlation of the TRF significantly below zero around 400ms post stimulus, it is also significantly below this reversed speech baseline. The next experiment used a different data set which showed that when you make the speech less intelligible through adding noise, the correlation disappears. Adding additional information such as a video of the speaker makes it easier to follow along for the subject and reintroduces the negative correlation, although a bit later than in the first experiment. In the final experiment a data-set was used where multiple story lines were spoken at the same time. When asked to focus on a single one, the correlation was similarly negative for the story line in focus and around 0 for the other.

A key weakness in the three experiments above is that each of them focus on showing evidence of semantic processing and the N400 component, as each shows, in a different way, that the negative correlation is present when the speech is intelligible and followed by the subject, and around 0 when not. No experiments on the strength of the negative correlation, the N400 effect, were actually done. This is surprising as the methodology includes surprisal values which are actually necessary for the N400 effect and not for the N400 component. The question then is whether simpler input vectors that only specify the location of the onset of words are not sufficient. A comparison between a simplified input vector and surpsal vectors would not only answer this question but also show possible evidence of the actual N400 effect. For if the simple vectors are indeed sufficient to show the N400 component yet a more negative correlation is observed in combination with the surprisal vectors, then that difference could be seen as evidence of the N400 effect.

The authors also noted that while there are many similarities to the negativity found in the TRF and the N400, they are hesitant in calling it such. Partly because their curves peak somewhat sooner than those seen in literature, but more importantly due to the differences between dissimilarity and the predictability based cloze task employed by the classic N400 studies. While this theoretical difference is important, we believe that there are more significant methodological problems with their dissimilarity measure. The dissimilarity value for a target word is calculated using the Pearson's correlation of its word2vec embeddings with the average of the embeddings of the words that came before it in the sentence. The averaging of the word vectors performed here is problematic in that there is no guarantee that the resulting average is on topic, let alone a good representation of the meaning of the words it's made of. Two completely different sentences on different topics could have the same or very similar average embeddings. Another criticism of this method is that it's unaware of the sentence structure or grammar. It's simply comparing words to each other (or averages). A word could be on topic compared to the other words but make no sense in the overall sentence. An example of this is the sentence ”I like sweet desserts therefore I am a pudding.” The word pudding is on topic compared to the other words but still surprising in this context.

It is surprising that, even taking into account these methodological shortcomings, a correlation was observed by Broderick et al. that shares so many similarities to the N400 component. This could indicate that, in line with more classic N400 component studies, the dissimilarity measure has no actual effect other than indicating the onset of a word.

3.1 Re-implementation

We replicated the methodology of Broderick et al. as closely as possible in order to take a critical look at the dissimilarity measure they introduced. Not only do we look at whether it’s necessary for finding evidence of semantic processing akin to the N400 component, but also whether it encodes enough information to capture evidence of the actual N400 effect. Both of these are done by adding additional random and static baselines for comparison.

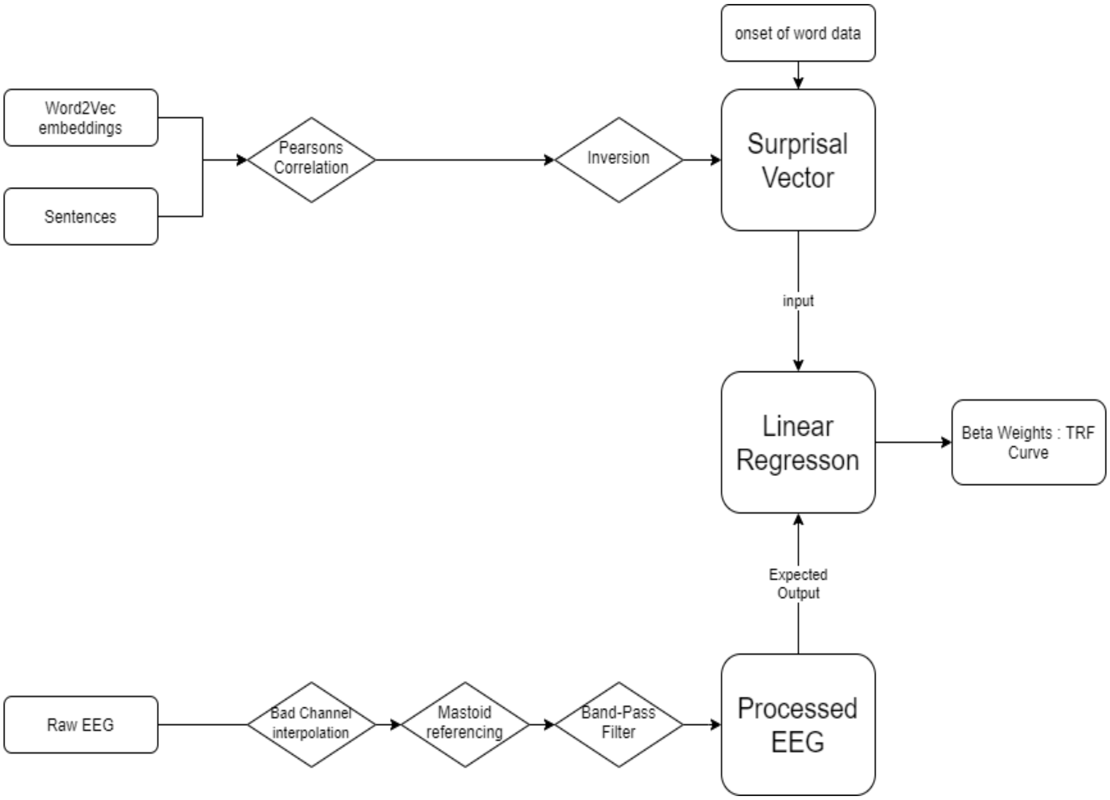


Figure 2: flow chart illustrating the original Broderick model setup. Boxes indicate processes and diamonds indicate information.

Figure 3 shows an overview of the methodology applied in the Broderick et al. paper.

There are two main components, the creation of the surprisal vector and the processing of the EEG. The Surprisal vector functions as the input to the linear regression model which optimises its weights such that the output generated is as close as possible to the processed EEG. These weights can be plotted to show the correlation of the input compared to the EEG for each time point in the 0 to 0.6 seconds time period post stimuli. This graph is known as the Temporal Response Function (TRF).

3.1.1 Data

The EEG Dataset used is the Natural speech dataset provided by the Broderick et al. paper. It consists of 19 subjects, 13 of which are male, who each listened to 20 runs of three minutes each of an audiobook of "The old men and the sea" by Ernest Hemingway. [2]

3.1.2 Dissimilarity based surprisal vector

The surprisal vector used by Broderick et al. is based on the similarity of words compared to their context. After removing the function words, for each word we look up the corresponding word2vec embedding and the average of the embeddings of the words that came before it in the sentence. The similarity of these two embeddings (target vs average embedding) are then calculated using Pearson's correlation. Subtracting this similarity from 1 gives the final surprisal value for the target word. As this is done for each word in each sentence in each run, the result is a 3D array with surprisal values.

The values created above made through dissimilarity calculations are used as an analogue for how surprising a word is given its context. The next step is to create the sparse surprisal vector by inputting these dissimilarity values at the spot in the vector that belong to the onset of the corresponding word in the audio book. As the vector itself has the exact same properties as the EEG data, the end result is 128hz vector, roughly 3 minutes long, that is 0 everywhere except for the time spots corresponding to the onset of a spoken word. These spots contain the surprisal value of that word.

3.1.3 Preprocessing the EEG

Prepossessing consists of three steps, two of which are identical to Brodericks approach: applying a band pass filter over the EEG and referencing it using the mastoids. The band pass filter is applied to remove all frequencies outside of the 1 to 8hz range as we are mostly interested in the low frequencies in the data. Referencing reduces the effect of head movements on the data. By looking at the two mastoids, electrodes that are placed on the EEG cap but not in contact with the scalp of the subject, we can find the

[2] A resource table is included in Appendix A with links to where to acquire this dataset and all other resources used in this paper

noise generated by movements of the head without any brain activity. The average of both the left and right mastoid are subtracted from the real EEG channels to remove this noise without touching the actual signal.

Where we differ is in the selection of bad channels. Broderick et al. manually selected which channels are considered to be bad. Not only is this tedious work, it also makes perfect reproduction impossible as these kinds of manual tasks depend heavily on the judgement of the researchers. For these reasons we have introduced an automatic system that finds the bad channels based upon the variance of the EEG signal in each channel over time. Note that this is a rough measure as there are multiple reasons why a channel could have a higher variance than another. We only select the channels with an intra channel variance higher than 100.000, as these can not be anything other than outliers. These bad channels are replaced with the spline interpolation of the surrounding channels

3.1.4 Learning: TRF

Now we can calculate the beta-weights which form the TRF through regularised linear regression. Usually in linear regression you train your weights on a provided training set of data that contains both stimulus and response with the goal of creating a model that can predict a response from a set of new stimuli. In our case we are not interested in predicting the EEG activity of new dissimilarity values but rather in the weights of the model as they encode the relationship between the EEG and dissimilarity value. The prediction task of the Linear regression can be mathematically expressed by:

$$r(t,n) = \sum_{d}^{D} w(d,n)s(t-d) + \varepsilon(t,n), \tag{1}$$

where:

$r(t,n)$ = EEG response at time t for channel n
$s(t)$ = stimulus (dissimilarity vector) at time t
$w(d,n)$ = unknown weight for channel n and delay d
t = seconds from the start discretized into 128hz
$\varepsilon(t,n)$ = Error of this time t and channel n
D = list of all delays in the recording period of 0 to 0.6 seconds in 128hz

This formula can be interpreted as follows. The EEG activity of e.g, the fourth second of channel ten ($r(4,10)$) is equal to the prediction of the model ($\hat{r}(4,10)$) + the residual activity unexplained by the model ($\varepsilon(4,10)$). The prediction ($\hat{r}$) itself is the weighted sum of the values of the previous 0.6 seconds on the dissimilarity vector. The weighting is done by the beta-weights ($w(d,n)$) which tells us how important the dissimilarity value of 0.4 seconds ago is compared to that of 0.2 seconds for example. Another way of looking at it is that the beta-weights w(d,n) are the filter of the convolution done on the dissimilarity vector, where each possible 0.6 seconds interval results in a predicted EEG activity of channel n for the last time point in that interval

TRF methodology visualised

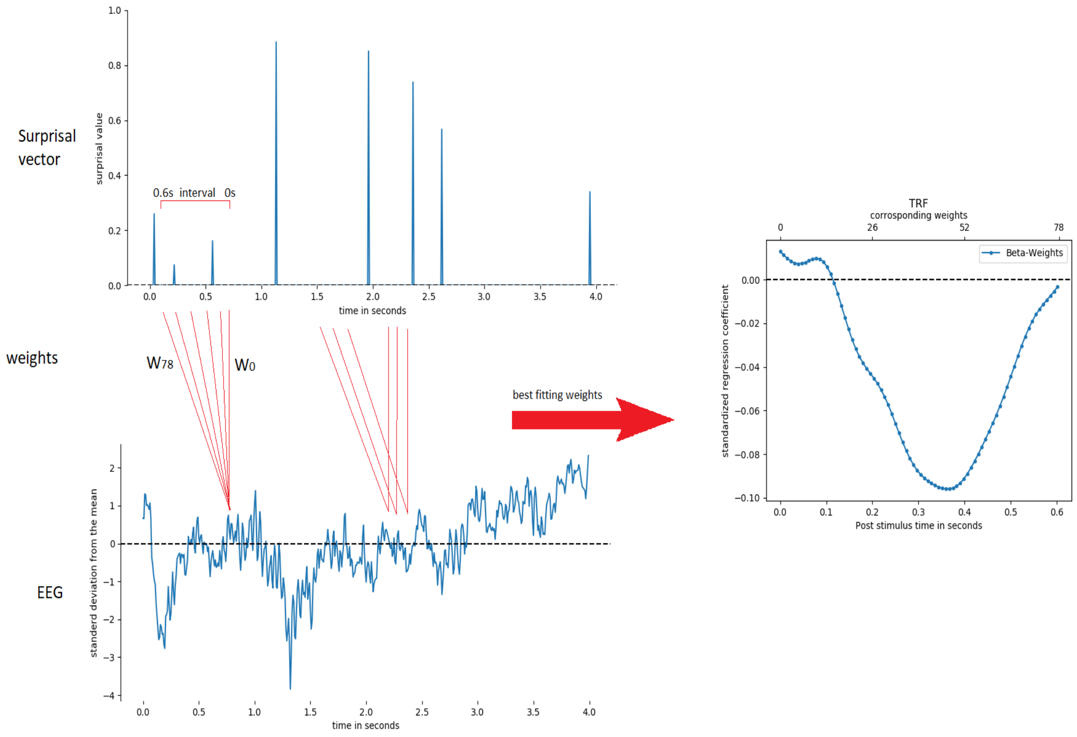
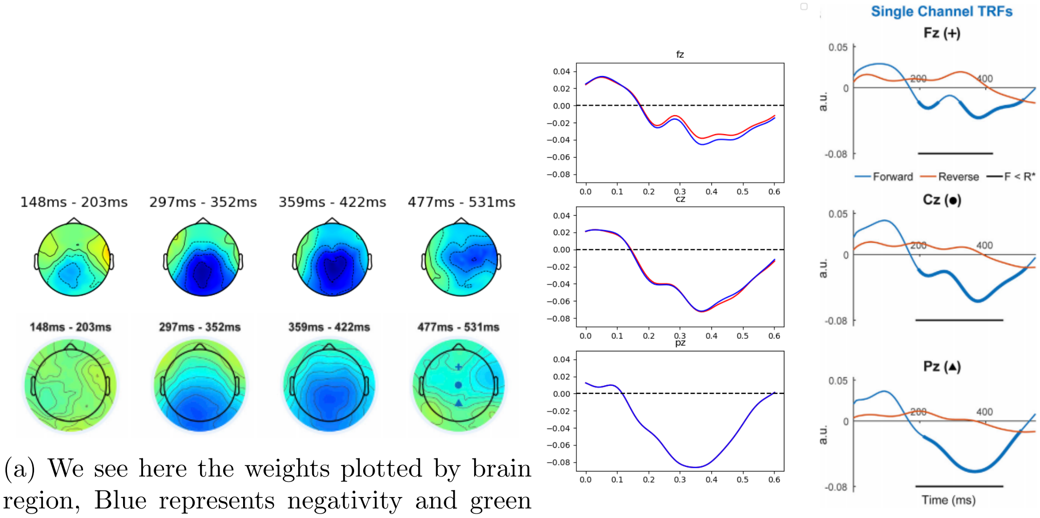


Figure 3: Shown here is a simplification of the process. Up top are shown a portion of the surprisal vector (SP) and the same portion of the processed EEG. The surprisal vector is a sparse vector where each non zero represents the onset of a word with its height equal to the surprisal value of that word. 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.

3.2 Results



(a) We see here the weights plotted by brain region, Blue represents negativity and green positive. Each plot is the average of a specific time period post stimulus. Up top are our reproduction results, below them are the results from the original Broderick paper. The bottom right image contains three symbols which correspond to the location of the Fz, Cz and Pz channels respectfully.

(b) To the left are our reproduction results. The blue line represents the reproduced Broderick results with automatic bad channel selection based on intra-channel variance. The red line represents the same method but with no bad channel interpolation at all. Note that in the Pz channel both lines are on top of each other. To the right are the results from the original paper. The blue lines are the original Broderick results we are reproducing. The orange line is their baseline made by playing the audiobook in reverse.

Figure 4a shows our results compared to the original Broderick et al. paper. The negativity is largely located in the same spatial region and the same temporal pattern is found by looking at the frontal (Fz), central (Cz) and parietal (Pz) midline channel TRFs in 4b. Our implementation differs in that bad channels are not picked by hand but instead done by picking the channels with extremely large intra channel variance instead. To show that any effect caused by this is minimal, the Fz, CZ, and Pz TRF's also show the same run without interpolating any bad channels.

In figure 4b we also see a glimpse of the baseline used in the original paper on the right. The orange line represents the TRF captured when the audiobook fragment was played in reverse to a portion of the same participants. The black lines underneath the graph represent the time periods where the blue line is significantly below the orange across all subjects. This shows that the N400 curve observed is related to semantic processing similarly to the N400 component. But what about the N400 effect?

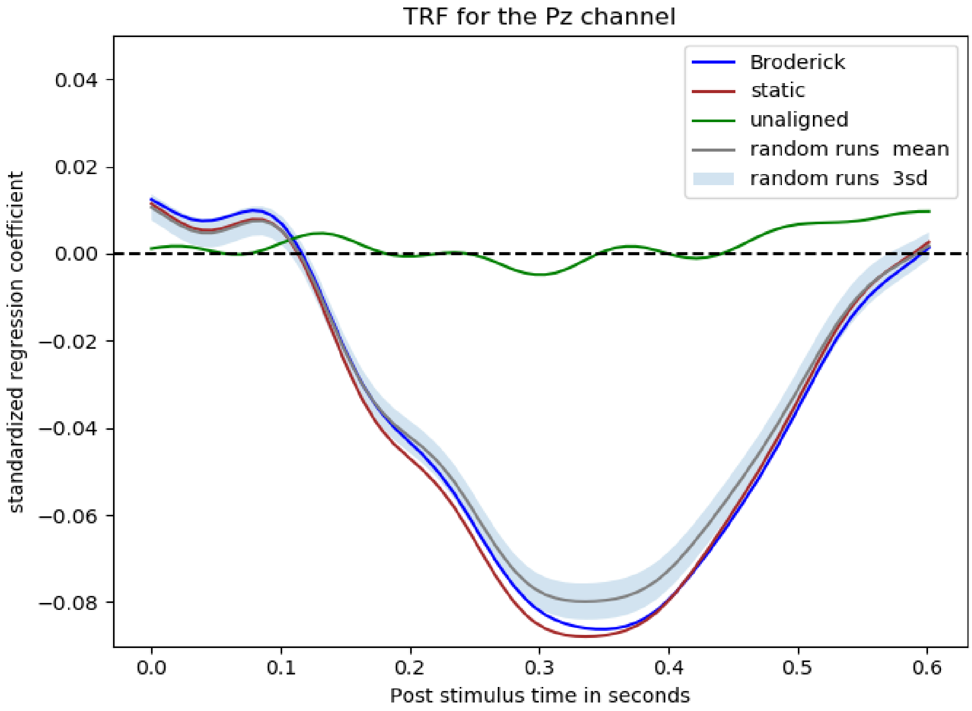


Figure 5: The TRF of the Pz channel with additional baselines for comparison.

To test this we have introduced additional baselines shown in figure 5. Here we see the TRF of the central-parietal scalp (the Pz channel). The unaligned baseline in green is made by putting surprisal values at random spots in the surprisal vector instead of at the onset of words. As a result, no measurable correlation was found. In essence this baseline can be seen as similar to the one Broderick et al. used in 4b, it shows that semantic processing needs to be done to see the characteristic N400 curve.

The gray line is the mean of 40 runs made with randomised dissimilarity values at the onset of words. Below it we see the brown baseline which is made by putting a one at the onset of each word instead of a dynamic number indicating how surprising that word is compared to its context. This surprisal vector is now 0 or 1 everywhere, with the 1s indicating the onset of words. While it looks promising to have the original Broderick model show stronger negative correlation than the randomised values and have its peak outside its 3 standard deviation range, as this suggests that that with correct surprisal values the model is capable of finding a more negative correlation, the static baseline seems to contradict this. Here we see an even more negative correlation in the relevant area between 0.3 - 0.5 second post stimulus

3.3 Discussion

The results so far indicate that the dissimilarity based surprisal values have no real effect, or even an adverse effect. While they contain some information as shown by their better than random performance, what they do contain is still mostly noise. In essence this suggests that the Broderick at all study could have been done by only indicating with a one where the onset of words are instead. None of their results would have likely changed for the worse without the surprisal calculations. These results make sense if you consider that their tests are made more akin to tests of the N400 component which is shown to be present for both surprising and non surprising words. This also means that no evidence of the N400 effect has been found, the question here is whether the problem lies in the methodology used for calculating surprisal or whether it's a limitation of the TRF based methodology.

4 Looking at the N400 effect through BERT

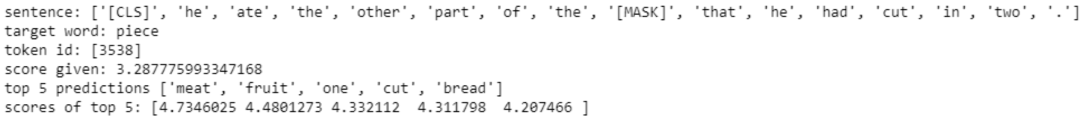
As a direction for improvement, Broderick et al. suggested moving towards a more predictability based surprisal vector, which is exactly what we have done. This means that we need a model that can predict how likely a word is to occur given its context. BERT was trained using the masked language task by taking inspiration of the same Cloze task that the original N400 study is based upon. This makes it more in line with classical N400 effect research which should result in a better approximation of human semantic processing and likely a stronger negative correlation with the EEG.

Besides the theoretical advantages of using surprisal like other more classic N400 experiments, this method also addresses the methodological problems of the previous implementation. No averaging is required as BERT directly looks at the whole sentence. A language model such as BERT is also completely aware of the sentence structure and word order. Theoretically, it would realise that human beings can normally not be puddings.

One problem in the use of BERT in this task is one of its main selling points: its bidirectional understanding of the sentence. BERT was made partly to address the difficulty of previous models to incorporate forward and backward looking models together. This was done by training on sentences where only the target word was masked and the words before and after the target in the sentence was accessible to BERT. This is problematic because in most cases, humans (and their brain) are only aware of the words they have heard so far. To look at the effect of this property of BERT on the results we have created multiple versions of the predictability based surprisal vectors. In half of these we have masked not only the target word, but also the words after it to simulate a more one directional approach. A problem that occurs with this approach is that BERT has a limited amount of information to base its predictions on, particularly when the target word occurs at the start of the sentence. To combat this, we have made additional versions for both forward and non forward looking BERT surprisal vectors that are provided 0 to 4 additional sentences that came before it as additional context for a total of ten variations. The naming scheme chosen for these models is as follows: " bx(LA)" with the x indicating how many sentences are provided as additional context followed by an LA for the look ahead models.

4.1 Implementation: Predictability based surprisal vectors

The goal here is to create a new surprisal vector of the same shape and magnitude as the previous dissimilarity one but this time based on predictability of the target word through BERT. As such the only difference is in how the surprisal is calculated, with all preprocessing of the EEG and the linear regression being the same. For each word in each sentence in each run, we mask the word itself and also the words after it if we are creating a non look ahead surprisal vector. This sentence is then given to a pretrained BERT model with 0 to 4 additional sentences as context. BERT provides a log likelihood for each word in its vocabulary at that spot, we simply select the one that corresponds to the original word. After doing this for all words we are left with a 3d array with all log likelihoods for each word in each sentence in each run. We normalise all these values to be between 0 and 1 through min-max feature scaling over the whole array. These values are then inverted by subtracting them from one and reshaped to surprisal vectors in the same way as the original vectors. Note that all surprisal vectors including the baselines are also normalized to have the same mean value as to make them comparable to each other.



```
sentence: ['[CLS]', 'he', 'ate', 'the', 'other', 'part', 'of', 'the', '[MASK]', 'that', 'he', 'had', 'cut', 'in', 'two', '.']
target word: piece
token id: [3538]
score given: 3.287775993347168
top 5 predictions ['meat', 'fruit', 'one', 'cut', 'bread']
scores of top 5: [4.7346025 4.4801273 4.332112  4.311798  4.207466 ]
```

Figure 6: An example of the forward looking Bert model with no additional sentences in action. We see in order: the sentence itself with the start token and a word masked. The word we want the log likelihood of. Its id in the total vocabulary. The log likelihood given to the target word by BERT (score). The 5 most likely words according to BERT. The likelihoods of the top 5.

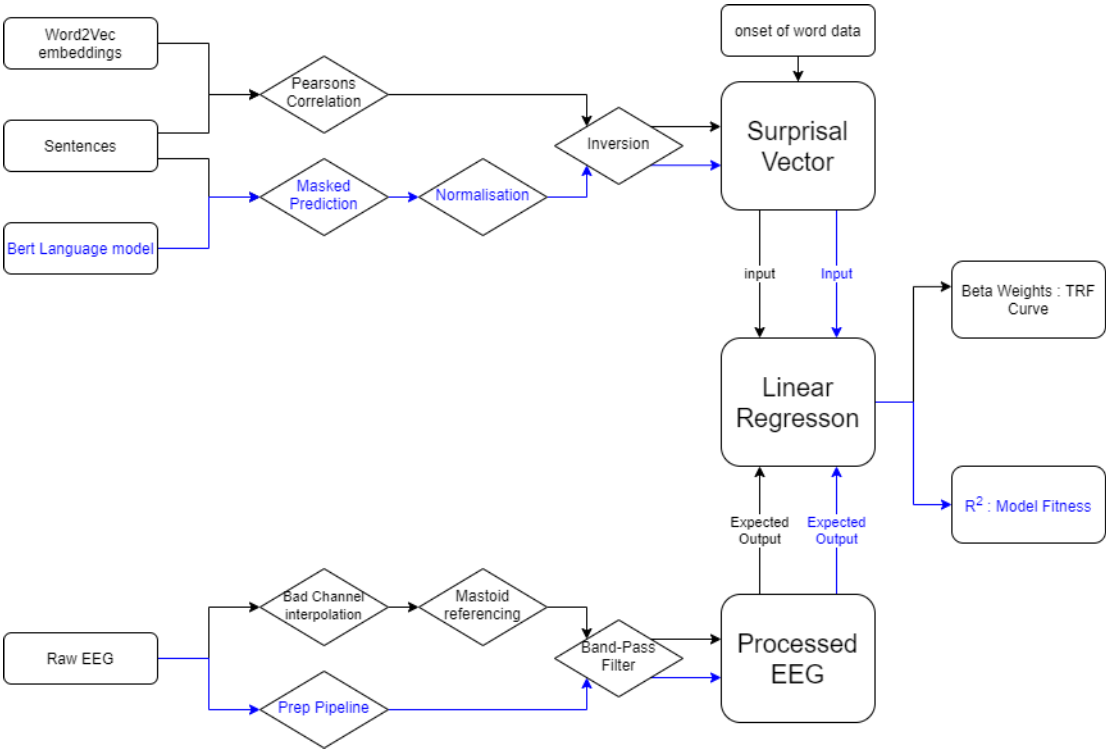


Figure 7: The expended flowchart of how the model works. Here additional blue boxes and diamonds have been added which represents additions we have introduced to the methodology. The additions include the new surprisal vectors created with BERT, the additional R^2 measure to show model fit and finally an additional pre-processing pipeline discussed later on.

4.2 Results

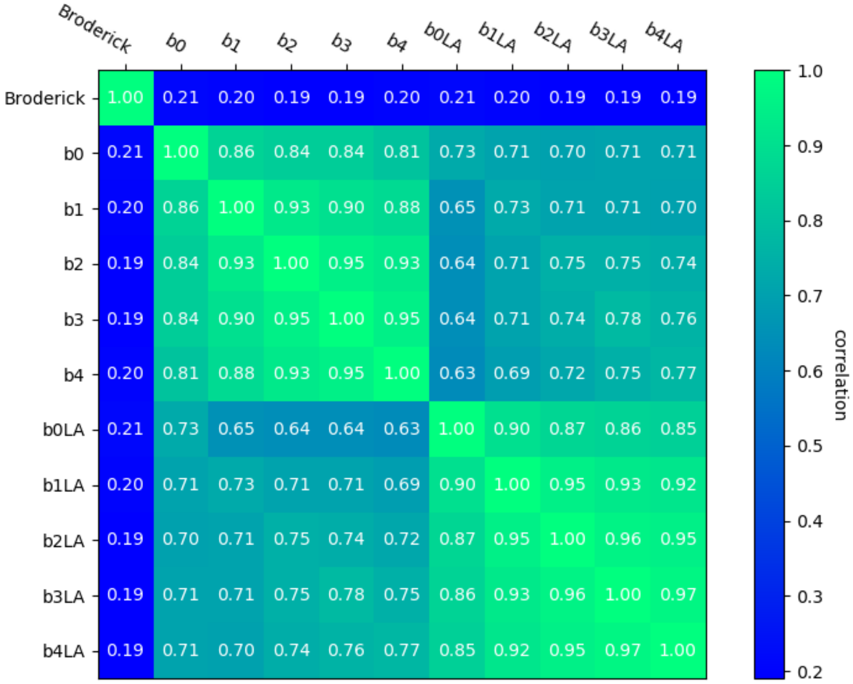


Figure 8: A heatmap showing the correlation between the different surprisal vectors.

The first thing we look at is how similar the new approach is to the old one, by using a correlation matrix over the surprisal vectors. Figure 8 shows this correlation matrix. There are two major observations here. As expected, the Broderick vector seems to be only loosely correlated to the BERT vectors, indicating that they largely disagree on which words are surprising. There is also a lesser disagreement between the BERT vectors which do look ahead versus those who do not which in turn suggests that they capture slightly different information.

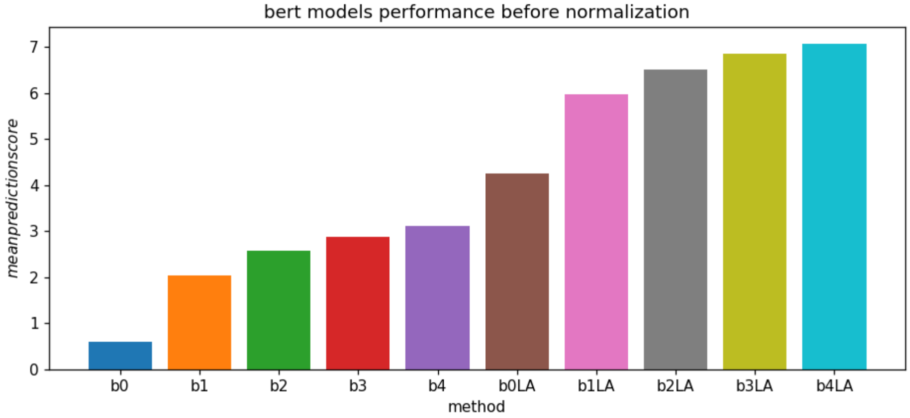


Figure 9: Mean prediction score of the different BERT models before applying normalisation. When you train BERT, it gives a score (the log-likelihood) based upon how likely each word in its vocabulary is on that spot. Here we see the mean of that score given to the target words.

To identify how the look ahead and non look ahead models differ, we take a look at the raw predictive power of each version. Figure 9 shows the mean log likelihood given by BERT for all words before normalisation. We see here that providing additional sentences as information improves the predictive power of both versions, but not to such a degree that the best non look ahead model competes with the worst look ahead one. This can likely be explained in part by the fact that BERT is trained on and for situations where it can see the complete sentence.

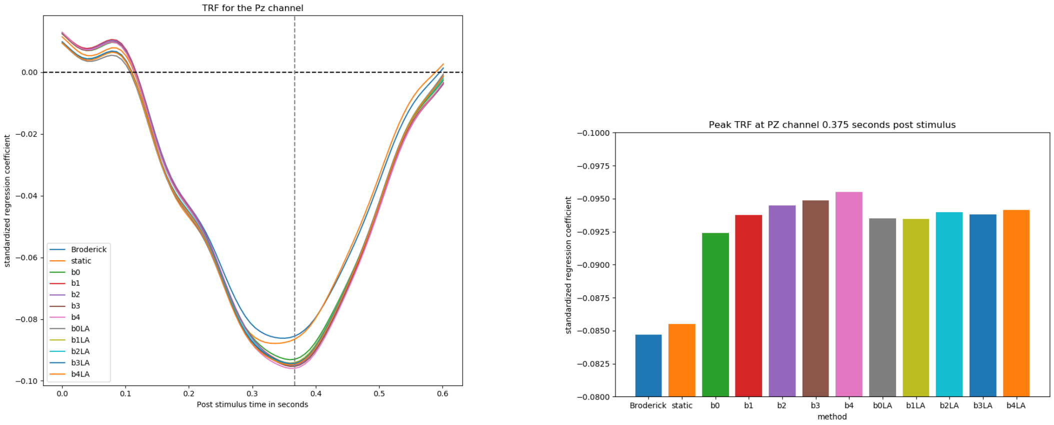


Figure 10: result of applying different versions of BERT similarity values compared to the Broderick version on the PZ channel. The right side shows the cross section highlighted by the gray vertical dashed line.

Above we see the main results of our paper, on the left side of figure 10 we see that irrelevant of which version is deployed, the model seems to be able to find a much more negative correlation between the EEG and the BERT surprisal values compared to the static baseline. Unlike the original Broderick version, the BERT based surprisal vectors do seem to encode actual information useful to the linear regression model. Additionally the peak negativity is later than in the BERT models compared to the original model.

Looking to the right we see that even though the look ahead models are more powerful when looking purely at how good of a predictor they are, the non look ahead models perform quite a bit better in this downstream task. Except for the first one which has no additional sentence as context, but this can be explained by the low amount of information provided to this version of BERT as it is prone to making large predictive mistakes, see figure 9.

4.3 Discussion

From the past three graphs we can conclude that the look ahead and non look ahead models capture different information from each other. The non look ahead models show more negative correlation likely because they function more similarly to humans. The look ahead models instead have partly future sight based superhuman prediction powers that do not help in correlating with human EEG. As such, even when their predictive power increases through a larger context, their correlation does not.

So we see here that the more theoretically and historically similar to the classic N400 effect prediction values are more negatively correlated than the original dissimilarity based surprisal values and more importantly, the static baseline. That their curve also peaks more closely to the classical N400 and that the more human-like implementation performs better than the more powerful one that is less similar to how humans function. All these together paint a clear picture that what we see here is likely to be the actual N400 effect described in neuroscience literature.

5 Model fit and preprocessing

Another part we want to explore in this research is the model fit of the TRF model. Considering the TRF methodology itself makes some problematic assumptions regarding the data it's working on. In particular the linear and time invariant assumptions are unlikely to be true when applied to the human brain. In our case there is no theoretical evidence that there is a linear relation between the input (surprisal) and the output (negative EEG deflection). It is also clear that the temporal dimension plays a big role here. The TRF treats each time point as an independent value with no direct relationship to the minutes around it. The EEG data is seen as a spreadsheet of data points in the regression with no understanding of its properties as a time series. Besides these assumptions, the TRF model is based upon simple regularized linear regression with only a single input while EEG data is known to be large, complex and noisy. Taking this into account, how good is the model fit? We look at this through plots showing the

R^2, a measure of how much signal to noise is captured by the model as a percentage from 0 to 1.

Additionally, we look at how critical some of the prepossessing steps are as we have seen in figure 4b that skipping on bad channel interpolation has minimal effect. This is likely because the normalisation performed on the EEG to create beta weights neuters the effect such bad channels have on the results.

We also applied a standardised method of preprocessing on the dataset called the PREP pipeline (Bigdely-Shamlo, Mullen, Kothe, Su, & Robbins, 2015). The PREP pipeline is made to standardize preprocessing for EEG across studies, with a focus on preparing it for machine learning tasks such as the one applied here. At its base it includes line noise removal, more complex bad channel detection and interpolation, and referencing to the true average. With the true average being the average of all channels after removing noisy/bad channels and other effects that might dominate it. We run the prep pipeline with the standard values recommended by the original authors.

5.1 Results

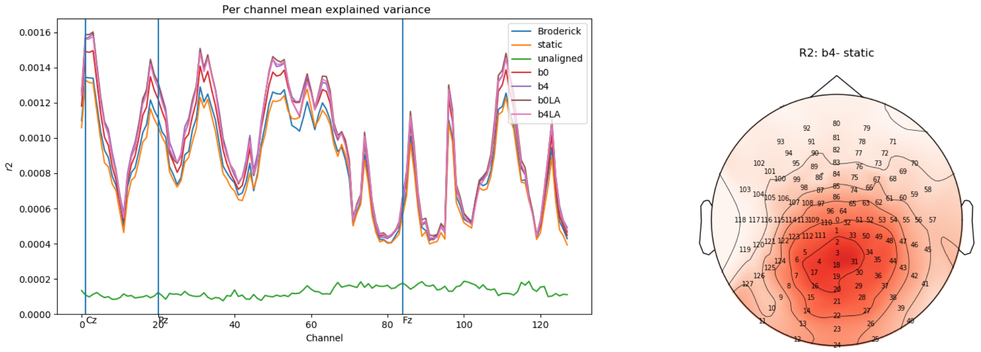


Figure 11: On the left, a curve plot of the R^2 of a verity of models with the individual channels on the X axis. On the right we see a topographic comparison between the b4 model and the static model made by subtracting the R^2 of each channel in the static model from those of the b4 model. The number corresponding to each of the channels are plotted in their corresponding locations.

Here we see the model fit through the R^2 of most of the models we have seen so far. The main takeaway here is the overall low R^2 for all the models. It is clear that the model is under fitting the data significantly. This is of course expected, as the EEG data is extremely large and noisy. What we do see is that the BERT models have a higher R^2 across the board compared the original Broderick model and the static baseline. On the right we see that these improvements come mainly from the regions associated with the N400 effect. As a sanity check we have included the unaligned version which is as expected incapable of finding any signal as it is trying to fit random data.

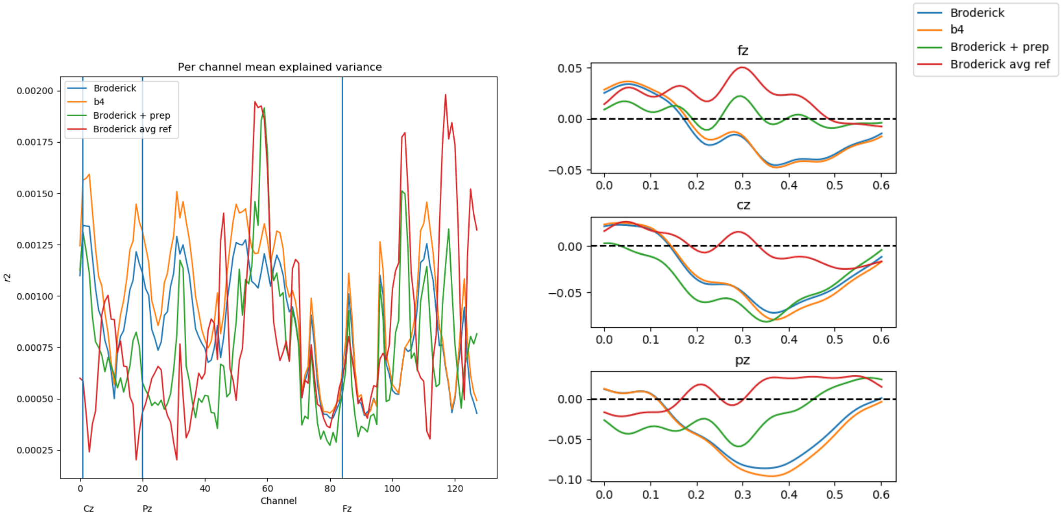


Figure 12: Comparison of the standard Broderick model with and without applying the PREP pipeline, a version of the Broderick model with average referencing instead of the normal mastoid referencing and the standard b4 model.

As shown by figure 12, the Prep model not only has a lower signal to noise ratio than the standard Broderick model, it also loses the characteristic shape associated with the N400. When looking at the standard model with average referencing instead of mastoid referencing, we see the same TRF and R^2 shape. It is therefore likely that what prevents PREP from working on the dataset is the referencing method applied as it's also based on average referencing. This is in line with classic literature, as the original N400 study defined the N400 as a deflection on the back of the brain compared to the mastoids. The R^2 of PREP is often above the other averaging method which suggests that the line noise removal component likely has a beneficial effect. Future research might want to try a version of PREP that applies mastoid referencing instead.

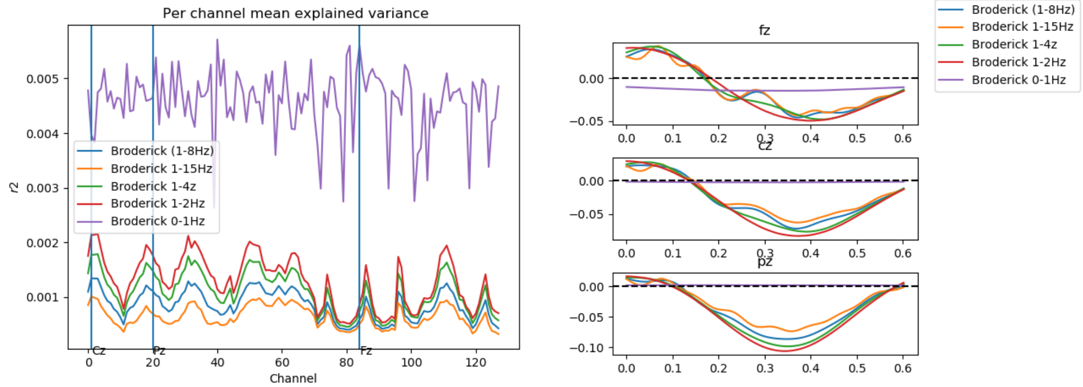


Figure 13: a look at the effect of the filter size upon the R^2 and the TRF of the Broderick model

Lastly we show how the model fit seems to improve when you narrow down the filter. This also seems to smooth out the TRF curve by creating a more convex shape. While this seems to be mostly beneficial, we are uncertain whether this means that the N400 can be mostly found in the lower frequencies or whether this is an artefact of how the model prefers lower frequencies perhaps due to its time invariant properties. The latter explanation is enforced by looking at the model fit of the 0 to 1 Hz range. We see a large increase in the model fit with no real N400 curve as the N400 is likely drowned out by other effects in this range. We also note that the smoothing effect provided by switching to the 1 to 2 Hz range is applied consistently across all the models seen in this paper and hence does not change any of the results shown so far in a significant way.

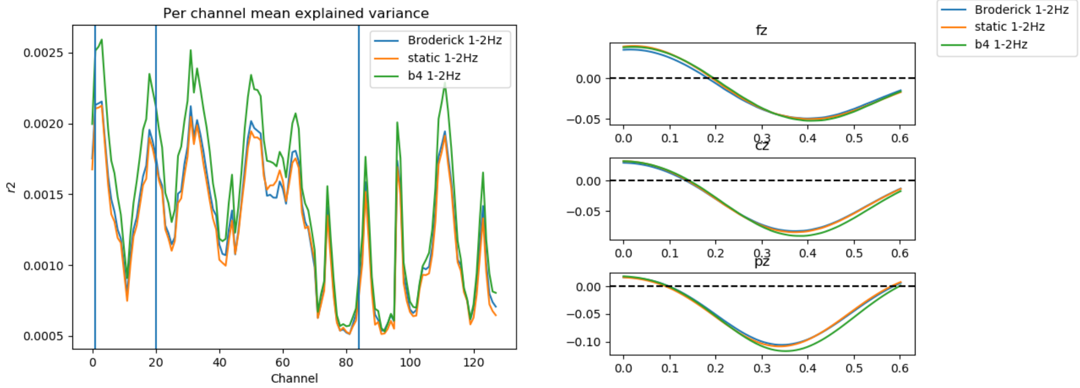


Figure 14: Comparison in the 1-2Hz range of the Broderick, static and best performing BERT model.

6 General discussion

The original N400 studies focused on their correlations with the Cloze tasks (Kutas & Hillyard, 1980). By using a NLP model that mimics this task such as BERT it is possible to find evidence of the N400 effect through the use of a much more relaxed experimental setup then those deployed in those classic studies. This evidence can be found by comparing the size of the negative correlation in the time period and location associated with the N400 to that of a static baseline, as this comparison shows a much more negative correlation between the EEG data and the BERT based surprisal vectors. We have also shown that even though the BERT models which look ahead in the sentence (and encode partly different information) have higher raw predictive performance, they perform worse in the downstream task as their correlation is less negative then the non look ahead models. This highlights that what we see is likely to be the actual N400 effect, as the more human-like model performs better at this task even if they are worse predictors. The N400 component can be found without the use of any linguistic information, as the static baseline alone is enough to invoke it. While this is inline with more classical N400 component research, we differ here from the original paper this one is based upon which used superficial dissimilarity based surprisal values to find them.

The model fit of all the models are quite low in general, they all have very low signal to noise ratios across the board. We can improve upon the signal to noise ratio somewhat by using a narrower band-pass filter as part of the prepossessing step. Not only does that increase the R^2 of the model, but it also smooths out the curves to a more convex shape. An even narrower band pass filter that only allows information in the 0-1Hz range significantly improves the R^2 but removes the characteristic curve associated with the N400. We note some uncertainty on whether these phenomena are a property of the N400 or a side effect of the inner workings of the TRF methodology.

6.1 Future Research

One avenue of exploration is the low signal to noise ratio of the models. We have tried to tweak the preprocessing of the EEG with some success but this issue can likely only be addressed by using a more complex model. The linear assumption is not present in neural network based models of which the recurrent variants are also well suited to time series as they take into account the temporal nature of the data. Ideally a LSTM or transformer of sorts would be used to understand the EEG, but this is unlikely to work well in practice due to the large amount of data required to make sense of the complex EEG. A more feasible approach could be to perform dimensional reduction through e.g. an autoencoder to construct a more compact representation of the EEG with minimal signal loss which can be fed into a neural network. The idea here would be to use an encoding format that takes into consideration the temporal and spatial nature of the EEG so that such a model would directly address both fundamental issues mentioned here of the TRF methodology.

Another direction would be to focus on more exploratory research. We have shown that there seems to be other bits of information captured by the TRF model that overpowers the N400 when we remove the band pass filter. Future research could try to find what exactly is captured, to determine whether it's just an artifact of the data or link it to a known or perhaps unknown brain property related to surprisal or dissimilarity. Unsupervised machine learning methods such as the aforementioned autoencoder could also be used to find more general patterns in the data that occur when being read an audio-book.

6.2 Conclusion

The original Broderick et al. paper showed evidence of semantic processing in continuous EEG recordings by looking at the TRF of how surprising a word is given its context compared against the EEG recording following the onset of that word. They found a negative correlation very similar to the N400 component but were hesitant to associate this negative correlation with it in part due to how their surprisal measure was based upon dissimilarity compared to the usual predictability based ones more commonly associated with the N400. We showed that this dissimilarity based measure is flawed and improved upon it by introducing a predictability based measure through the use of the BERT language model. This new approach is not only capable of finding a negative correlation between a word being spoken and its EEG like the original model, but also shows that a better surprisal model has a more negative correlation. This suggests that more surprising words have stronger deflection similar to the N400 effect.

Additionally we note that the use of a narrower band pass filter focused on the lower frequencies smooths out the TRF curve and increases the signal to noise ratio of the model. Whether this is a product of the N400 itself or the inner workings of the TRF model is uncertain. In general we find that the fit of the model itself is very low, likely due to underfitting on the complexity of the EEG data. We believe future research should therefore move towards better computation models, preferably those who do not share the linear time invariant properties of the TRF.

References

Abnar, S., Ahmed, R., Mijnheer, M., & Zuidema, W. (2017). Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. *arXiv preprint arXiv:1711.09285*.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 238–247).

Barrett, S. E., & Rugg, M. D. (1990). Event-related potentials and the semantic matching of pictures. *Brain and cognition*, *14* (2), 201–212.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The prep pipeline: standardized preprocessing for large-scale eeg analysis. *Frontiers in neuroinformatics*, *9*, 16.

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, *16* (13), 4207–4221.

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.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, *10*, 604.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry Jr, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20* (4), 400–409.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... others (2013). Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, *7*, 267.

Harris, Z. S. (1954). Distributional structure. *Word*, *10* (2-3), 146–162.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9* (8), 1735–1780.

Holcomb, P. J., & Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. *Language and cognitive processes*, *5* (4), 281–312.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing* (Vol. 1, pp. 181–184).

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

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Kutas, M., Neville, H. J., & Holcomb, P. J. (1987). A preliminary comparison of the n400 response to semantic anomalies during reading, listening and signing. *Electroencephalography and Clinical Neurophysiology Supplement*, *39*, 325–330.

Marmarelis, V. Z. (2004). *Nonlinear dynamic modeling of physiological systems* (Vol. 10). John Wiley & Sons.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008a). Predicting human brain activity associated with the meanings of nouns. *science*, *320*(5880), 1191–1195.

Rugg, M. D. (1985). The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology*, *22*(6), 642–647.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high-and low-frequency words. *Memory & cognition*, *18*(4), 367–379.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2020). Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

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).

Appendices

A Resources table

Resources table

resource	source	url
EEG recordings	Dryad, (Broderick et al., 2018)	`https://datadryad.org/stash/dataset/doi:10.5061/dryad.070jc`
word2vec embeddings	(Baroni et al., 2014)	`http://marcobaroni.org/composes/semantic-vectors.html`
Bert Language model	(Devlin et al., 2018)	`https://github.com/Meelfy/pytorch\_pretrained\_BERT`

Table 1: A list of the resources used, their source and a download link.

B Codebase

The codebase used in this paper can be found in `https://github.com/Rasyan/Thesis`. This includes the recreation of the Broderick et al. methodology as well as the additions discussed in this paper. All code is presented as a Ipython Notebook and commented. There are three main files included:

1. One which shows how to create the different surprisal vectors.
2. Another which calculates the TRF weights from each different surprisal vector and each method of prepossessing.
3. The last notebook contains the code for creating the different graphs of this paper using those weights.

To simplify reproduction, the commands to create all the surprisal vectors and corresponding TRFs used in this paper are provided at the end of each file.