Meaning in brains and machines: Internal activation update in large-scale language model partially reflects the N400 brain potential

Alma Lindborg (lindborg@uni-potsdam.de

Milena Rabovsky (milena.rabovsky@uni-potsdam.de)
Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24-25,
14476, Potsdam, Germany

Abstract

The N400 brain potential has been used as a neural correlate of meaning-related processing in the brain, but its underlying computational mechanism is still not well understood. Although efforts to model the N400 as an update of a probabilistic representation of meaning have been promising, the limited scope of earlier models has restricted experiments to highly simplified sentences. Here, we expand modelling of the N400 to naturalistic sentences using a large-scale, state-of-the-art deep learning language model (GPT-2). We investigate the correspondence between updates in the internal state of the model and the N400 in one quantitative experiment and four qualitative experiments. Our findings suggest that activation updates in the model correspond to several N400 effects, but cannot account for all of them.

Keywords: sentence comprehension; event-related potentials; N400; deep neural networks; language models

Introduction

Language ultimately aims to convey meaning. The brain potential that is most commonly used to investigate the processing of meaning in language is the N400 component of the event related brain potential (ERP), a negative deflection over centro-parietal electrodes peaking around 400 ms after the presentation of a potentially meaningful stimulus. N400 amplitudes have been shown to be modulated by a large number of lexical-semantic variables, being larger for sentences with semantic violations (*'I take my coffee with cream and dog'* as compared to *'sugar'*), low cloze as compared to high cloze probability continuations, earlier rather than later positions of a word in a sentence, target words presented after an unrelated as compared to a related prime (*table – dog* rather than *cat – dog*), and words that generally occur with lower as compared to higher lexical frequency, among many others (see Kutas & Federmeier, 2011, for review). Despite its widespread use and even though it seems relatively clear that the N400 is somehow related to meaning processing, the functional basis of the N400 continues to be actively debated. Recently, Rabovsky, Hansen, and McClelland (2018) and Rabovsky (2020) simulated a broad range of overall 17 distinct empirically observed N400 effects using the Sentence Gestalt (SG) model (St. John & McClelland, 1990), a neural network model of sentence comprehension that maps from sequentially incoming words to the situation or event described by a sentence. N400 amplitudes were simulated as the magnitude of change of an internal hidden layer activation state that implicitly represents expected sentence meaning. The observed correspondence between N400 amplitudes and this activation change was taken to suggest that N400 amplitudes reflect the change of a probabilistic representation of meaning corresponding to an implicit semantic prediction error (see also Rabovsky & McRae, 2014).

A limitation of this previous modelling work was that the SG model was trained on a small synthetic corpus and thus could not be presented with the same naturalistic stimuli presented in empirical experiments. There are various ways to address this limitation. One way is to scale up the SG model (Lopopolo & Rabovsky, 2021). The other way, which we pursue here, is to ask whether the change of a probabilistic representation of meaning as implemented in the SG model can be similarly captured by state-of-the-art large-scale language models, which bear some family resemblance to the SG model, such as e.g., GPT-2 (Radford et al., 2019).

The GPT-2, released by OpenAI in 2019, is a state-of-the-art deep neural network language model trained for next-word prediction (Radford et al., 2019). Like most current state-of-the-art language models, the GPT-2 is based on the Transformer architecture (Vaswani et al., 2017). Although not explicitly trained to perform other tasks than next-word prediction, the GPT-2 performs surprisingly well in a variety of language tasks such as text summarisation, machine translation and question answering (Radford et al., 2019). This makes it an interesting candidate for modelling sentence comprehension in a more naturalistic setting than the limited universe of the small scale SG model.

The GPT-2 relies on masked self-attention, whereby the model remembers all previous inputs in a text segment and dynamically assign attention to words which are deemed to be important in processing the current word. This means that even distant previous words can attract the model's attention when it processes complex material with long-range dependencies. Although this unreasonably large memory buffer means that Transformer models process language in a less biologically plausible fashion compared to recurrent models such as the SG, recent studies suggest that they outperform recurrent models in predicting N400 amplitudes (Merkx & Frank, 2020) as well as psychometric measures such as reading times (Wilcox, Gauthier, Hu, Qian, & Levy, 2020; Merkx & Frank, 2020). However, these studies focus on the output of the models (such as model-derived lexical surprisal), and say little about how meaning is represented internally by Trans-

former models. Another strand of recent work suggests that activation patterns in Transformer models such as the GPT-2 can accurately predict activity in the human language system (Schrimpf et al., 2020; Caucheteux & King, 2020). These results are particularly surprising given the relatively low biological plausibility of Transformer models compared to recurrent models, and raise the question of whether these models form on-line semantic representations similar to those of the human brain. This question has not yet been directly addressed.

For the purposes of modelling processing of meaning in general, and the N400 specifically, it is clear that the SG model and the GPT-2 are two very different creatures. Whereas the SG model is prompted to map a sentence to the corresponding described situation or event, and thus learns by linking language to the world, the GPT-2's training is language internal and the model is only implicitly engaged in interpreting the meaning of the sentence insofar as it is needed for the prediction of its continuation. Nonetheless, its (mostly) coherent continuations of stories in response to prompts and its performance in a variety of language comprehension tasks suggest that it does implicitly represent meaning. Furthermore, several recent studies have argued that it is exactly this principle of prediction that underpins the correspondences between Transformer models and brain activity (Caucheteux & King, 2020; Schrimpf et al., 2020; Heilbron, Armeni, Schoffelen, Hagoort, & de Lange, 2020). Thus, the prediction-driven spontaneous interpretation the GPT-2 is engaging in may be an interesting model of human language processing, but it is most likely carried out within a subspace of the model (for example, certain layers of the deep neural network), and may be associated with certain computational stages. For example, in analogy with evidence of hierarchical processing within deep neural networks for image classification (Khaligh-Razavi & Kriegeskorte, 2014), one may hypothesise that integrated semantic representations may arise at deeper layers of the GPT-2. Finally, the notion of 'update' may have different meanings in the two models due to the differences in how they process text. Although they both read a sentence from left to right and interpret each word in the context of previous words, the SG model forms a single, integrated representation of the sentence which is updated incrementally, whereas the GPT-2 retains access to previous states by the attention mechanism. Thus, in the GPT-2, the interpretation of a sentence may not be fully contained in its last state, but rather distributed over many of its states. This means that the GPT-2 could employ other strategies than incrementally updating a single representation of meaning. Although recent work suggests similarities between recurrent models and transformers (Katharopoulos, Vyas, Pappas, & Fleuret, 2020), it remains to be investigated whether activation updates in the GPT-2 show similarly interpretable patterns as those of the SG model.

In this study, we test whether the correspondence between the N400 and updates in the probabilistic representation of meaning in the SG model (Rabovsky et al., 2018) can be extended to the GPT-2. If this is the case, representations of meaning in the GPT-2 may provide insight into those of the human brain. Our investigation is two-pronged. In a quantitative experiment, we compare updates in the GPT-2 to electrophysiological responses in an EEG experiment containing no explicit experimental manipulations of the N400. Here, we quantitatively evaluate the evidence for a correspondence between the N400 potential and updates in internal states of the GPT-2 on naturalistic sentence stimuli. Additionally, in four qualitative experiments, we investigate the effect of explicit experimental manipulations on the internal dynamics of the GPT-2.

Experiments

Experiments were run using the pre-trained GPT-2 model `gpt2-large` publicly available from the Huggingface `transformers` library (Wolf et al., 2020). This model implementation contains an embedding layer followed by a stack of 36 Transformer decoder modules, which each has 1280 units in its output layer. These activations are used as inputs to the next decoder module, and finally used to estimate a probability distribution over possible next words. Although the model contains more units – notably the attention modules – we focus our analyses on the decoder output activations. We investigate whether the finding by Rabovsky et al. (2018) that the amplitude of the N400 correlates with the size of the update in the representation of sentence meaning generalises to the GPT-2. In analogy with their study, we define the size in network update $u(n)$ at word position n as

$$u(n) = \sum_{i=1}^{D} |a(n)_i - a(n-1)_i| = \|a(n) - a(n-1)\|_1 \qquad (1)$$

where $a_i(n)$ is the activation of unit i in some D-dimensional layer, to the presentation of word n. In our N400 experiments, n will be the position of the target word, for example *sugar* in the stimulus sentence *'I take my coffee with cream and sugar'*, which will be compared to the activation at word $n-1$, i.e. *and*. We calculate the update at each decoder output layer separately, resulting in one update measure for each of the 36 decoders. As the GPT-2 may split words into multiple tokens each associated with an internal state, we defined the update at word w_t as the change in activation from the first token of w_t and the last token of the previous word w_{t-1}.

We tested the correspondence between the N400 and activation update in the GPT-2, defined in Equation (1), using two complementary approaches. In the quantitative experiment, we compared the GPT-2 update to the N400 using experimental data collected by Frank and colleagues (Frank, Otten, Galli, & Vigliocco, 2015, 2013). We presented the same stimuli to the GPT-2 as to the subjects in the electrophysiological study and investigated whether the update significantly predicted the N400 amplitude of the subjects. Moreover, we

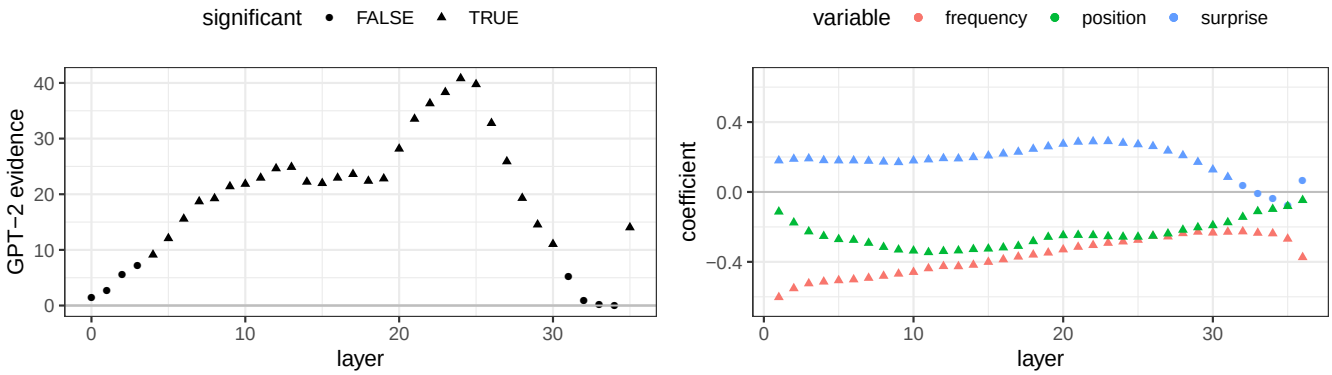

Figure 1: Results of the quantitative experiment, plotted by layer in the GPT-2. The left panel shows evidence for the GPT-2 update predicting the N400, indicated by the χ^2 test statistic. Significant χ^2 values are indicated with triangles. The right panel shows the influence of three lexical and sentence level variables on the GPT-2 update. Significant coefficients, marked with triangles, have the same sign as the corresponding regression coefficients for the N400, indicating that the GPT-2 update and the N400 are similarly influenced by these variables.

investigated whether lexical-semantic variables that naturally vary across the stimuli such as word frequency, sentence position and lexical surprisal, have a similar influence on the GPT-2 update as on the N400.

In the qualitative part of the study, we investigated whether the GPT-2 update is sensitive to explicit experimental manipulations of semantic congruence, expectancy, role reversals and priming. These manipulations have all produced clear effects on the N400 in earlier electrophysiological experiments. Here, in the absence of EEG data we tested whether experimental conditions differed in the same direction as reported in previous experiments. The stimuli in the qualitative experiments may be less familiar to the GPT-2 than those in the quantitative experiment, as the GPT-2 is trained on texts from the internet, reflecting the statistical patterns in online texts. Thus, although the 'unnatural' stimuli in such experiments may mean that they are less representative of language comprehension under naturalistic conditions, they are a test of the generality of the model's behaviour. The priming experiment lies furthest away from the GPT-2's training data, probing the model on word pairs rather than the sentence stimuli it is trained to parse, in contrast to previous sentence-based priming experiments on transformers (Misra, Ettinger, & Rayz, 2020). Ideally, a model correlate of the N400 would give reliable and clear effects also for stimuli which are relatively unfamiliar to the model.

Quantitative experiment

In the quantitative experiment, we investigated whether updates of the GPT-2 can predict the N400 amplitude in response to naturalistic sentences. Additionally, we investigated whether the GPT-2 updates are similarly modulated by three naturally varying lexical and sentence level variables as the N400: sentence position, lexical frequency, and lexical surprisal. The N400 is known to decrease through the course of a sentence, with lower amplitudes for words presented late in the sentence (Van Petten & Kutas, 1991). Moreover, infrequent words yield a larger N400 compared to frequent words (Van Petten & Kutas, 1990), as do words which are surprising given the previous words in the sentence (Frank et al., 2013).

We used a publicly available EEG dataset, collected for a previous study conducted by Frank et al. (2013). The experiment material consisted of 205 sentences, which did not contain semantic or syntactic violations, or other types of unnatural language. Event-related potentials were collected from 24 subjects, who read the sentences word by word in a pseudo-random order. Lexical frequency was calculated as each word's log-transformed word frequency in the British National Corpus, and surprisal was estimated using an n-gram model, following the original study (Frank et al., 2015). For more details on the data acquisition and EEG pre-processing, see Frank et al. (2013, 2015).

Investigating first whether the GPT-2 update can predict the N400, we computed the GPT-2 updates in response to the experimental stimuli, at each of the model's 36 decoder output layers. For each set of GPT-2 updates representing a layer of the model, we fit a linear mixed effects model predicting the N400 of the experimental subjects from the updates. In addition to the GPT-2 updates, we included a fixed effect of the ERP baseline amplitude, which was not subtracted from the N400 component following the original study (Frank et al., 2015). Random intercepts for each subject and lexical item were furthermore included in the model. Subsequently, each of these 36 linear mixed effects models were compared to a base model including only the fixed effect of ERP baseline and the random effects of subject and lexical item. The evidence for the effect of the GPT-2 update was estimated for each layer using a likelihood ratio test between the model including the GPT-2 update and the base model. The evidence was deemed significant if the χ^2 statistic satisfied $p < 0.05$

after Bonferroni-Holm correction for multiple comparisons. Results of the model comparison are displayed in the left panel of Figure 1. The model evidence for the GPT-2 update was significant in 31 of the model's 36 layers, with insignificant layers all being close to either the input layer or the output layer of the GPT-2. The evidence appeared to be largest (as indicated by a large χ^2-value) in the deep intermediate layers $21-25$.

Secondly, investigating the influence of the three lexical-semantic variables on the GPT-2 update and the N400, we estimated the contribution of each variable to the GPT-2 update and the N400, respectively, and compared the direction of the effects. To this end, we fit a linear mixed effects model at each layer of the GPT-2, predicting the GPT-2 update at each item from its sentence position, lexical frequency frequency and surprisal, with a random intercept for lexical item. Regression coefficients were extracted at each layer and deemed significant if their t-value satisfied $p < 0.05$ after correction for multiple comparisons. Similarly, we fit a linear mixed-effects model predicting subjects' N400 amplitude from the same independent variables, additionally including a fixed effect of the baseline ERP amplitude and random intercepts for lexical item and subject. As the N400 is a negative ERP component, we multiplied the amplitudes by -1 in order to facilitate comparison with the GPT-2 updates (which are non-negative). The mixed effects model of the N400 revealed a positive coefficient for surprisal ($\beta = 0.064$, $t = 7.49$, $p < 0.0001$) and a negative coefficient for sentence position ($\beta = -0.058$, $t = -10.42$, $p < 0.0001$). Additionally, there was a negative regression coefficient for lexical frequency ($\beta = -0.007$, $t = -0.514$, $p > 0.05$), but this was not significant in the current dataset, contrary to other N400 studies (Van Petten & Kutas, 1990).

The regression coefficients for the lexical and sentence level variables on the GPT-2 updates are plotted in the right panel of Figure 1. Apart from the 5 deepest layers, the variables significantly affect the GPT-2 update in the same direction as the N400: regression coefficients for surprisal are positive, whereas those of lexical frequency and sentence position are negative.

Qualitative experiments

In the qualitative experiments, we followed classical N400 experimental paradigms, recording the response to a target word w_t presented in a context $w_1, \ldots, w_{t-1}$. We conducted four experiments, listed in Table 1. Experimental conditions were compared statistically by one-sided paired t-tests in the direction of the hypothesis, using a significance level of $\alpha = 0.05$ corrected for multiple comparisons by the Bonferroni-Holm method.

Table 1: Qualitative N400 Experiments

	Experiment	**Hypothesis**
1.	Semantic violations	violation > congruent
2.	Cloze probability	unexpected > expected
3.	Reversal anomalies	incongruent > reversal $\geq$ congruent
4.	Priming	unrelated > related

In **Experiment 1**, we tested the effect of semantic violations (such as *'I take my coffee with cream and dog'*) on the GPT-2 activation updates. In line with the well-known result from Kutas and Hillyard (1980) on the N400, we predicted that semantically incongruent sentence continuations elicit a larger update compared to congruent continuations (such as *'I take my coffee with cream and sugar'*). This hypothesis was tested using a set of 350 congruent and 350 incongruent sentences, collected by Valderrama, Beach, Sharma, Appaiah-Konganda, and Schmidt (2020).

In **Experiment 2**, we tested the effect of the cloze probability on target words, as demonstrated by Kutas and Hillyard (1984). Here we compared the update for expected endings (*'The children went outside to play'*) to improbable, but not strictly incorrect endings (*'The children went outside to talk'*). We used a set of 498 sentences with high-cloze endings constructed by Block and Baldwin (2010), with manually added low-cloze completions of the same sentences. We predicted a larger N400 (and thus a larger update in the GPT-2) for low cloze compared to high cloze endings.

In **Experiment 3**, we tested whether reversal anomalies produced similar effects on the GPT-2 update as on the N400. In reversal anomalies, such as *'For breakfast the eggs would only eat...'*, despite the implausibility of *eat* in this context, only a small increase in the N400 is observed compared to congruent sentences such as *'For breakfast, the boys would only eat...'* (Kuperberg, Sitnikova, Caplan, & Holcomb, 2003). Thus, we tested whether the GPT-2 update is smaller for reversal and congruent conditions compared to a fully incongruent condition (*'For breakfast, the boys would only plant...'*). Moreover, we tested whether the reversal condition produces a larger update in the GPT-2 compared to the congruent condition. We used a set of 180 sentences in congruent, incongruent and reversal anomaly conditions in this experiment, constructed by Kuperberg et al. (2003) [1].

In **Experiment 4**, we tested whether the GPT-2 network update could simulate the N400 reduction for semantically (eg. *'school – university'*) or associatively (eg. *'school – teacher'*) primed words compared to unrelated word pairs (Koivisto & Revonsuo, 2001). We selected the set of 1000 most strongly associated word pairs from the English Small World of Words project (De Deyne, Navarro, Perfors, Brys-

[1] The stimuli, kindly provided by the authors upon request, overlapped with those of the 2003 study but were not completely identical with them.

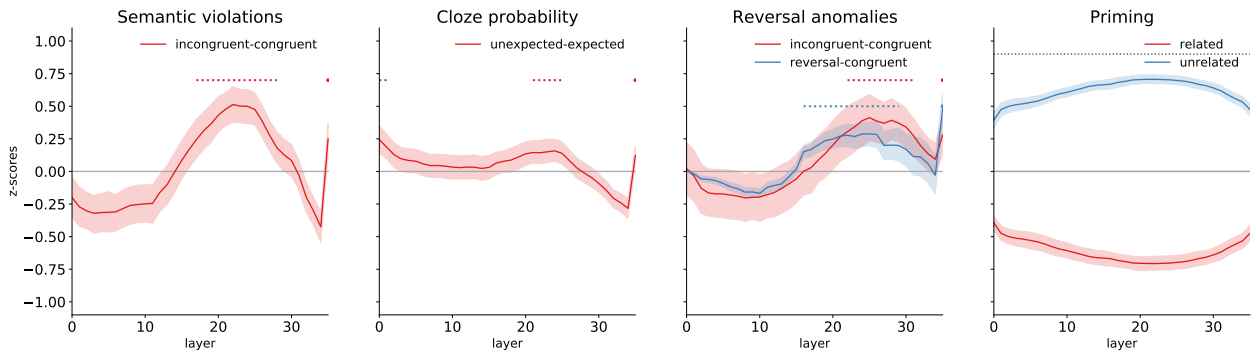

Figure 2: Results of the qualitative experiments, plotted by layer in the GPT-2. The solid lines represent the mean difference in update between conditions, with error bands representing a 95% confidence interval. In the Priming experiment, each condition is plotted separately, instead of their difference. The dotted lines indicate layers where significant effects in line with the hypotheses were found.

baert, & Storms, 2019) and used the first and second words as prime and target words, respectively. We compared the primed condition to an unrelated condition, containing the same set of targets with scrambled primes. Scrambling was done prior to tokenization, in order to avoid that words split up into multiple tokens were separated.

Results of the qualitative experiments are displayed in Figure 2. In Experiments $1-2$, we find the predicted differences (incongruent $>$ congruent in the semantic violations experiment, unexpected $>$ expected in the cloze probability experiment) localised mainly to the deeper intermediate layers of the model. In Experiment 3, effects are similarly located in the deeper intermediate layers of the GPT-2, but results are more mixed. Incongruent targets elicited a larger GPT-2 update than congruent targets, in line with the hypothesis – however, contrary to our prediction there was no significant difference between the incongruent condition and the reversal condition. Finally, in the priming experiment, the model's update is larger for unrelated target words as compared to primed, in line with the hypothesis. In contrast to the other experiments, the effect was significant in all layers of the GPT-2, indicating that word pair associations are encoded already at initial processing stages of the model.

Discussion

In this study, we investigated whether internal state changes in a deep learning language model – the GPT-2 – can be used to model on-line representations of meaning in human language comprehension, as measured by the N400 potential. Prompted by earlier research using the SG model (Rabovsky, 2020; Rabovsky et al., 2018), we used the update in internal representation as a potential correlate of the N400.

In the quantitative experiment, we investigated whether the GPT-2 update could predict N400 amplitudes from a reading experiment using naturalistic sentences. Here, we found significant evidence for the GPT-2 update predicting the N400 in all but the outer (shallowest and deepest) layers of the model (see Figure 1). However, this does not imply that all layers of the GPT-2 are necessarily engaged in meaning-related processing. Rather, as we found in our investigation into the effect of lexical-semantic variables on the GPT-2 update, early stages of processing are dominated by effects of lexical frequency and sentence position. As these variables also influence the N400, it is not surprising that even shallow layers of the GPT-2 can predict the N400 to some extent. However, the effects at the shallow layers may follow relatively trivially from the architecture and training objective of the GPT-2. First, prior probabilities for single words should be well-represented in a statistical model for next word prediction such as the GPT-2. Second, as Transformer models explicitly encode the position of each word token in the text input (Vaswani et al., 2017), it is plausible that the GPT-2 has learned to grade the importance granted to a word by its positional embedding. On the other hand, in humans sentence position effects on N400 amplitudes are presumably not due to direct position encoding, but rather due to increasing predictability of later as compared to earlier parts of the sentence.

The layers for which the evidence is highest are located deeper within the network, coinciding with an increase in the influence of surprisal on the updates. These layers seem to provide the best candidates for an N400 correlate within the GPT-2. A notable difference between the GPT-2 update and the N400 is the effect of lexical frequency on each measure. Whereas lexical frequency has the numerically largest effect on the GPT-2 update of all three variables, its effect on the N400 amplitudes in the corresponding EEG experiment was non-significant. Although a significant effect of lexical frequency on the N400 has been established in other studies (cf. Van Petten & Kutas, 1990), it is possible that the GPT-2 may be more focused on lexical frequency than human readers, due to its training on next word prediction. Further study is needed in order to ascertain whether this discrepancy between the model and human subjects is limited to the current EEG dataset or an indication of a general pattern.

In the second part of our investigation, we investigated the effect of specific experimental manipulations on the GPT-2 and tested whether experimental conditions differed in the same direction as the N400. Manipulating semantic congruence and expectation, respectively, we found significant N400-like responses in the deeper intermediate layers of the GPT-2 (see Figure 2). Here, we found that a larger update was elicited for semantic violations compared to congruent target words, and for unexpected compared to expected target words. Interestingly, these effects were observed at layers overlapping with those that most strongly predicted the N400 in our quantitative experiment.

In the reversal anomaly experiment, we expected a larger update for incongruent targets compared to the congruent and reversal condition. This prediction was not fulfilled, as we found no significant difference between the incongruent and reversal conditions. Thus, whereas the role reversal is not reflected at the N400 processing stage in humans, the GPT-2 seems to process these inputs differently. In this sense they resemble measures of surprisal, which have also been shown to capture N400 amplitudes quite well in general, but fail in some specific situations such as reversal anomalies (Rabovsky et al., 2018). Here, semantic update in the SG model is more in line with empirical N400 data, possibly because the SG model is not trained on next word prediction but rather is trained to estimate sentence meaning based on both syntactic and plausibility based constraints such as word order and event probability; see (Rabovsky et al., 2018) for discussion.

Finally, in the priming experiment, we found that related word pairs induced a smaller update in the GPT-2 compared to unrelated pairs, at all layers. This more wide-spread effect could potentially arise because the association between words is already represented at the stage of word embedding, and – it seems – preserved as the information is passed through the deep neural network.

In summary, we have found partial evidence for similarities in on-line semantic updates in the GPT-2 and the human brain, as indicated by correspondences between GPT-2 network updates and the N400 brain potential. In our quantitative experiment, we found that the GPT-2 update predicted N400 amplitudes from a reading experiment, with strongest evidence at deeper intermediate layers. Moreover, in our qualitative experiments we found modulations of semantic congruence and expectancy overlapping with the quantitative N400 effects, and finally we found a wide-spread effect of priming. These findings are notable, given that the GPT-2 (in contrast to the SG model) is not explicitly trained to estimate sentence meaning, nor is it architecturally constrained to incremental processing. Thus, the effects we found at deep intermediate layers of the GPT-2 suggest that incrementally updated semantic representations may to some extent be an emergent property of prediction-based language processing, even in the absence of the memory constraints imposed in recurrent neural networks. However, the absence of difference between the incongruent and reversal sentences in the reversal anomaly experiment suggests differences between the GPT-2 update and the N400, pointing to possible limitations to the validity of such a model as an analogy for how humans understand language. All in all, our results suggest that the Transformer architecture may not provide a perfect model for the neural process underlying the N400, but may well be useful in modelling certain types of meaning-related processes at a level of comprehension performance currently unmatched by more biologically plausible deep learning language models.

References

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670.

Caucheteux, C., & King, J.-R. (2020). *Language processing in brains and deep neural networks: computational convergence and its limits* (preprint). bioRxiv. doi: 10.1101/2020.07.03.186288

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. , 11.

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2020). *A hierarchy of linguistic predictions during natural language comprehension* (preprint). bioRxiv. doi: 10.1101/2020.12.03.410399

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020, August). Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. *arXiv:2006.16236 [cs, stat]*. (arXiv: 2006.16236)

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915. doi: 10.1371/journal.pcbi.1003915

Koivisto, M., & Revonsuo, A. (2001). Cognitive representations underlying the N400 priming effect. *Cognitive Brain Research*, *12*(3), 487–490.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129.

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647.

Kutas, M., & Hillyard, S. (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.

Lopopolo, A., & Rabovsky, M. (2021). Predicting the n400 erp component using the sentence gestalt model trained on a large scale corpus. In *Proceedings of the 43rd annual meeting of the cognitive science society.*

Merkx, D., & Frank, S. L. (2020). Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv: 2005.09471*.

Misra, K., Ettinger, A., & Rayz, J. (2020). Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4625–4635). Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.415

Rabovsky, M. (2020). Change in a probabilistic representation of meaning can account for N400 effects on articles: A neural network model. *Neuropsychologia*, *143*, 107466.

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), 693–705.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*(1), 68–89.

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

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2020). *The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing* (preprint). bioRxiv. doi: 10.1101/2020.06.26.174482

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*(1-2), 217–257.

Valderrama, J. T., Beach, E. F., Sharma, M., Appaiah-Konganda, S., & Schmidt, E. (2020). Design and evaluation of the effectiveness of a corpus of congruent and incongruent English sentences for the study of event related potentials. *International Journal of Audiology*, 1–8.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequencyinevent-related brain-potentials. *Memory & Cognition*, *18*(4), 380–393.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition*, *19*(1), 95–112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention Is All You Need. *arXiv: 1706.03762*.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *arXiv: 2006.01912*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv: 1910.03771*.