Computational Psycholinguistics — Assignment 2

Daan Brugmans (S1080742)

I. INTRODUCTION

This report is the realization of the Assignment 2 project for the Radboud University course Computational Psycholinguistics. For this assignment, students must investigate whether the gradients computed from a recurrent neural network correlate with measured P600 component activity from a controlled experiment. The reasoning behind this assignment is that recent research ([1], [2]) has shown that the P600 component may be the backpropagation of prediction errors in the human language system. Since neural language models also backpropagate their prediction errors using gradients, there may exist similarities between the language error backpropagation of human and artificial neural language systems. This report contains the findings found by me for the Assignment 2 project.

The relevant code for this assignment can be found at the following URL: https://github.com/daanbrugmans/ru-computational-psycholinguistics-23-24/tree/main/assignment-2/code.

II. RELATED WORK

For my realization of Assignment 2, I have chosen a controlled experiment from [3], specifically their first experiment. Kim and Osterhout researched P600 activity in anomalous sentences. When they released their research, it was generally accepted that the P600 component was solely responsible for language error related to syntax, while the N400 was responsible for language error related to semantics. In their work, Kim and Osterhout show that the P600 is still activated in sentences that are syntactically unambiguous, but have illogical semantics as a result, however. The general structure of such sentences was one where the two agents of a sentence are swapped, maintaining a certain "theme" and a valid syntax, while going against semantics that are considered regular. An example of this is the sentence "*The hearty meal was devouring the kids*", where the syntax is unambiguous, but does imply a meaning that goes against what is usually expected of the combination of the sentence's agents and theme, which would be "*The kids were devouring the hearty meal*". The authors name such sentences **(Attraction) Violations**.

Kim and Osterhout note that Attraction Violations, especially the phrase up and including the critical verb ("*The hearty meal was devouring*"), while syntactically unambiguous, can be interpreted to be either syntactically or semantically incorrect. If a reader interprets the violation as syntactically invalid, then they likely think that the critical verb is in the wrong tense, and changing it would return the sentence to a meaning that aligns with expectations ("*The hearty meal was devoured*"). If a reader interprets the violation as semantically
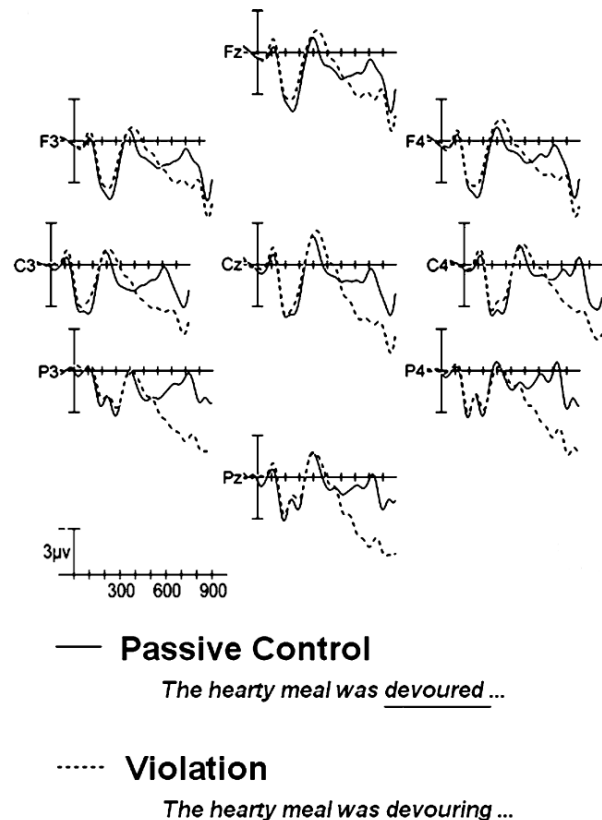

Fig. 1: ERP Brain Activity when reading Violation vs. Control sentences, from [3].

invalid, then they likely think that the agent is wrong, and, in the case of the full sentence, that the agents should be flipped ("*The kids were devouring*"). This means that an Attraction Violation can cause a syntactically unambiguous sentence to be considered syntactically invalid due to its semantics. On the basis of this phenomenon, Kim and Osterhout researched if sentences with Attraction Violations can trigger a P600 effect in humans. Although P600 effects were typically taken to be responsible for purely syntactic errors at that time, syntactically unambiguous Attraction Violation sentences, which can be considered to be syntactically erroneous on the basis of the sentences meaning, may thus elicit a P600 response based on a sentence's unexpected semantics.

Kim and Osterhout conduct an experiment where participants' N400 and P600 components are read while reading both Attraction Violation sentences ("*The hearty meal was devouring*"), and Control Sentences where the critical verb of the sentence is altered as to let the sentence be both syntactically and semantically coherent with its agents and

theme ("*The hearty meal was devoured*"). Their results show that, on the critical verb, the P600 component is much more pronounced on Attraction Violation sentences than on the Control sentences. They also showed that the N400 component is only slightly present when Attraction Violation sentences are read, meaning that the error stemming from the sentence is mostly processed by the P600. These findings then imply that the P600 is not responsible for syntactic error alone, and that a degree of semantic error is also processed in the P600. A visualization of these findings can be seen in figure 1.

III. METHODOLOGY

All the code I have written can be found in the Jupyter Notebook called `main.ipynb`, which should thusly contain all results also shown in this report. It can be found at the following URL: https://github.com/daanbrugmans/ru-computational-psycholinguistics-23-24/blob/main/assignment-2/code/main.ipynb I have placed the code in the `get_predictions.py` file in a function called `get_predictions`, so that I can easily call this code from the notebook.

Kim and Osterhout ([3]) provide their full set of Attraction Violation sentences and (Passive) Control sentences in their paper. I have copied the full sets of these sentences up until and including the critical verb. I have stored them as text files called `violations.txt` for the Attraction Violation sentences and `control.txt` for the Control sentences in a folder called `data`. The notebook processes these sentences further so that they can be used by the RNNs by lowercasing all words and separating genitive markers ('s) from consonants. Furthermore, after the surprisals and gradients have been calculated, I remove sentences that do not have surprisal and/or gradient information, and remove sentence-initial words from sentences that do.

My results are a collection of scatterplots that visualize the relationship between the surprisal and gradients of the models. Every plot also contains a quadratic function fitted to the data to show the trend of the relationship between the variables. I have chosen for a quadratic function as opposed to a linear function, as I found that this fits the data better. Included in every plot is also the Pearson Correlation Coefficient *r* between variables. Since surprisal values and gradients are calculated for multiple models, I provide a plot for every model. For every pair of variables, I also provide a plot of how the Pearson Correlation Coefficient changes as the model is trained on more data.

Before producing results, I set a universal seed for Python itself, NumPy, and PyTorch of `3131`. I do this in order to improve the reproducibility of my results. This is why I also set some settings for CUDA regarding PyTorch's randomness when computing on GPU.

IV. RESULTS

The full results can be found in the appendices, where the graphs for every model are shown for all four comparisons:
- Violation vs. Control Surprisal
- Violation vs. Control Gradients

Fig. 2: Surprisal vs. Gradients on Violations for an RNN trained on 300,000 sentences.

Trained Sentences Count	*r*
10,000	0.065
30,000	-0.094
100,000	-0.600
300,000	-0.601
1,000,000	-0.577
3,000,000	-0.431
10,000,000	-0.252
30,000,000	0.149
epoch1	0.282

TABLE I: Correlation Coefficients between Model Surprisal and Gradients on Violation Sentences by the model's train data size.

- Violation Surprisal vs. Control Surprisal
- Violation Gradients vs. Control Gradients

A. Comparing Surprisal vs Gradients

Figure 3 shows the change in Pearson Correlation Coefficient between a model's surprisal and gradients on sentences with an Attraction Violation as the model has been trained on more data. Figure 4 shows the same information, but for Control sentences. Tables I and II show the exact correlation coefficient values from these plots. Finally, figure 2 shows an example of a model's surprisals against its gradients, specifically the Recurrent Neural Network trained after 300,000 sentences.

B. Comparing Conditions

Figure 5 shows the change in Pearson Correlation Coefficient between a model's surprisal on sentences with an Attraction Violations versus Control sentences as the model has been trained on more data. Figure 6 shows the same information, but for the model's gradients. Tables III and
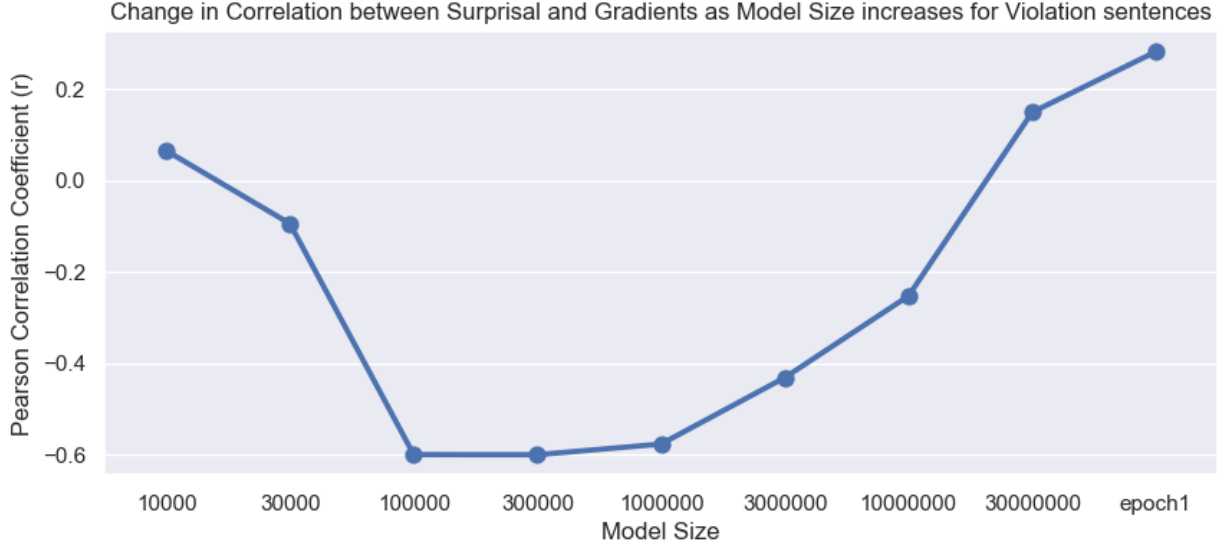
Fig. 3: The change in Correlation between Model Surprisal and Gradients on Violation Sentences as the model is trained on increasingly more data.
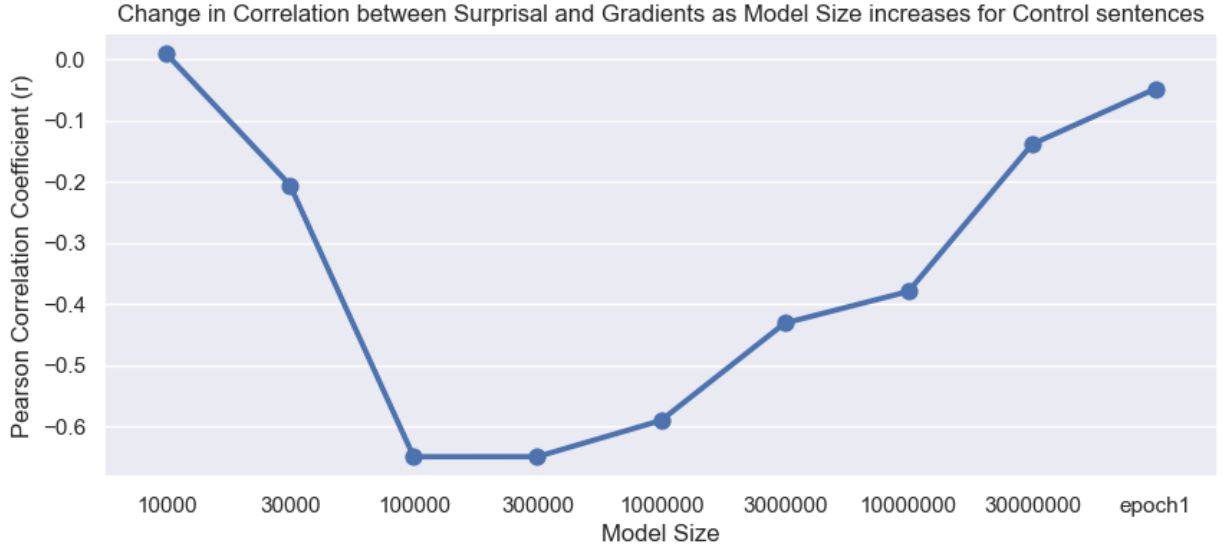

Fig. 4: The change in Correlation between Model Surprisal and Gradients on Control Sentences as the model is trained on increasingly more data.

Trained Sentences Count	r
10,000	0.009
30,000	-0.205
100,000	-0.650
300,000	-0.650
1,000,000	-0.591
3,000,000	-0.431
10,000,000	-0.380
30,000,000	-0.139
epoch1	-0.048

TABLE II: Correlation Coefficients between Model Surprisal and Gradients on Control Sentences by the model's train data size.

IV show the exact correlation coefficient values from these plots. Additionally, figure 7 shows the correlations between surprisals and gradients on the two difference conditions for the fully trained model.

V. DISCUSSION

A. Correlations between Surprisal and Gradients

The data in figures 3 and 4 show an interesting trend in the predictive power of a model's surprisal for its gradients: the strength of the correlation between the variables gets stronger as the model has trained on more data. This strength reaches a peak of -0.6 to -0.65 when the model is trained on 300,000 sentences, after which, the correlation strength starts to move into the opposite direction; it returns to zero and grows from negative to positive. I expect that the increasing strength of
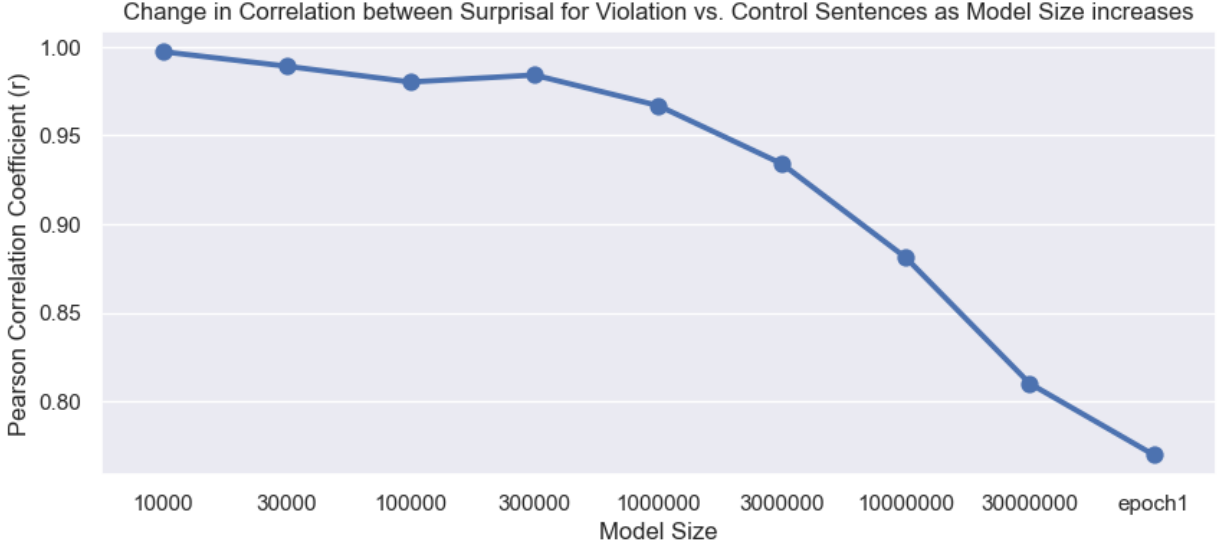
Fig. 5: The change in Correlation of Model Surprisal between Violation Sentences and Control Sentences as the model is trained on increasingly more data.
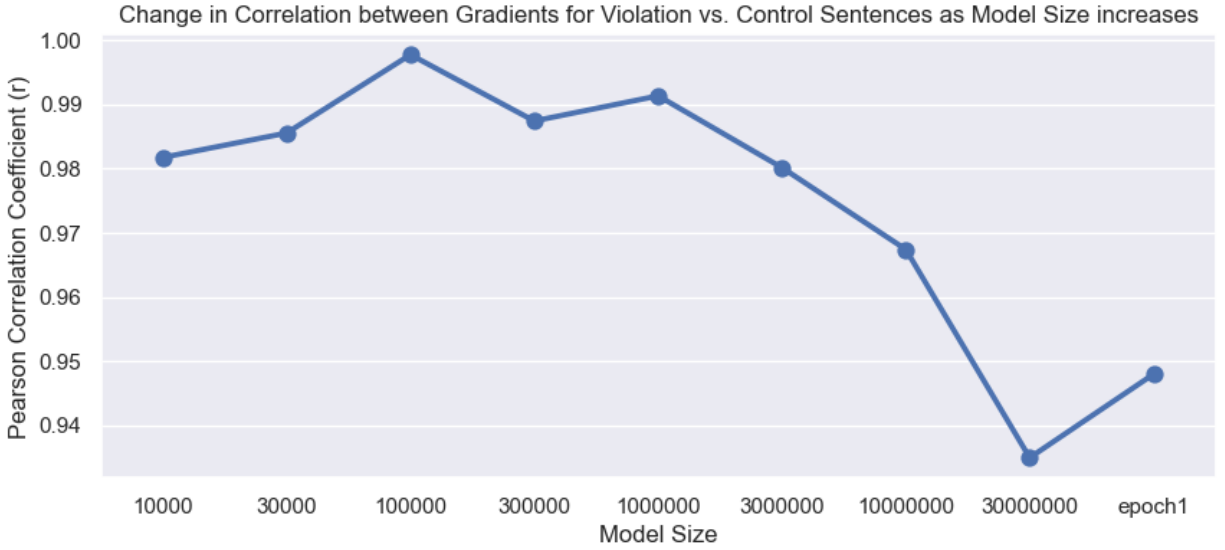

Fig. 6: The change in Correlation of Model Gradients between Violation Sentences and Control Sentences as the model is trained on increasingly more data.

Trained Sentences Count	r
10,000	0.998
30,000	0.990
100,000	0.980
300,000	0.984
1,000,000	0.967
3,000,000	0.934
10,000,000	0.881
30,000,000	0.810
epoch1	0.770

TABLE III: Correlation Coefficients between Model Surprisals of Violation Sentences and Control Sentences by the model's train data size.

Trained Sentences Count	r
10,000	0.981
30,000	0.986
100,000	0.998
300,000	0.987
1,000,000	0.991
3,000,000	0.980
10,000,000	0.967
30,000,000	0.935
epoch1	0.947

TABLE IV: Correlation Coefficients between Model Gradients of Violation Sentences and Control Sentences by the model's train data size.

the correlation until 300,000 sentences can be explained by the

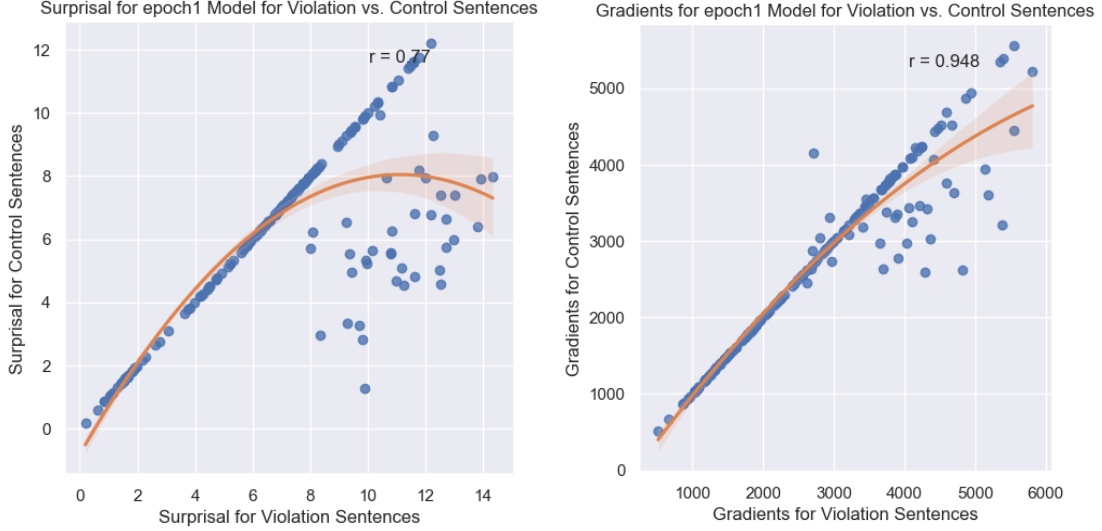learning process of the model. As the model has learned more,

Fig. 7: Plots of the correlation between the fully trained (epoch1) model's surprisals (left) and gradients (right) on Violation Sentences vs. Control Sentences.

its gradients become a better representation of its surprisal, because it has gained an understanding of the data it has learned. When it has reached this depth of understanding, words that it expects backpropagate with small losses, while words that it is surprised by backpropagate with large losses. As the model continues to learn, the surprisal starts to become a worse predictor of the gradients. I expect this is due to overfitting: the model doesn't learn from being surprised by words anymore, so the backpropagated losses represent other information that it is learning. Since these losses represent more than just surprisal, they become less correlated with surprisal. Based on this explanation, we will assume that the model's gradients are best explained by its surprisals and vice versa for the model trained on 300,00 sentences. Figure 2 shows the exact way in which the relationship between surprisals and gradients develops: higher surprisals are more common than lower surprisals, and as the size of the surprisal increases, the size of the gradients tends to decrease.

B. Correlations between Conditions

Figures 5 and 6 show the change between sentence conditions (Violation vs. Control) for a model's surprisal/gradients as the model is trained on more sentence data. These correlations stand for how similar the surprisals and gradients are between conditions: the higher the correlation, the more similar they are. A lower correlation between conditions then implies a greater difference in surprisal/gradients between conditions. This disparity between conditions says something about the effect strength of the anomalies in the Violation sentences: lower correlations imply that the syntactic/semantic anomalies have a bigger effect on the model's surprisal/gradients.

The trend in figure 5 is clear: as the model has trained on more sentence data, the correlation decreases, and does so at an accelerating rate. This would imply that, as the model is trained on more data, it should be more surprised about the critical verbs in Violation sentences. The left plot of figure

7 shows this in practice: the diagonal line in the plot is the collection of words preceding the critical verb, while the points off this line are the critical verbs. It can be seen that these critical verbs have higher surprisals in the Violation sentences than in the Control sentences. Thus, as a model has learned on more data, it becomes more surprised at unexpected words.

The trend of decreasing correlations across conditions does not seem to hold to the same degree for gradients, however. Figure 6 shows that there exists only a slight decrease in correlation as the model is trained on more sentence data. The right plot of figure 7 shows the details: the diagonal line stands for words preceding the critical verb, while the points off this line are the critical verbs. It can be seen that the critical verbs tend to elicit a greater backpropagation gradient in Violation sentences than Control sentences, although this is not always the case.

C. Correlations with ERP Components

VI. CONCLUSIONS

REFERENCES

[1] Hartmut Fitz and Franklin Chang. Language erps reflect learning through prediction error propagation. *Cognitive Psychology*, 111:15–52, 2019.
[2] Stefan L Frank. Neural language model gradients predict event-related brain potentials, Jan 2024.
[3] Albert Kim and Lee Osterhout. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2):205–225, 2005.

APPENDIX