A closer look at N400/P600 role reversal effects in the Retrieval-Integration computational model

Allyson Ettinger[1], **Tal Linzen**[3], **Colin Phillips**[1], **Philip Resnik**[1,2]
[1]Department of Linguistics, [2]Institute for Advanced Computer Studies
University of Maryland, College Park, MD
[3]Department of Cognitive Science, Johns Hopkins University, Baltimore, MD
{aetting,colin,resnik}@umd.edu, tal.linzen@jhu.edu

Abstract

We replicate the results of Brouwer et al. (2017) in simulating divergent N400 and P600 effects in semantic role reversals, and we examine the factors contributing to the success of the simulation. We find that the model's performance depends on specific assumptions about training data, which complicate interpretation of the model's success.

1 Introduction

The N400 and P600 event-related potential (ERP) components are instrumental in investigations of human language processing—and for understanding the mechanisms underlying these responses, phenomena that produce divergent effects between the components are particularly informative. One such phenomenon is that of semantic role reversals (e.g., "he forgot which waitress the customer served") which have been found across a variety of languages to induce P600 effects, but not N400 effects (e.g., Hoeks et al., 2004; Chow et al., 2016).

While there are multiple verbal theories of the N400 and P600 (Kos et al., 2010; Bornkessel-Schlesewsky and Schlesewsky, 2008; Kuperberg, 2007; Kim and Osterhout, 2005; Van Herten et al., 2005), to our knowledge only one computational model aims to simulate both components, successfully capturing their divergence on role reversals (Brouwer et al., 2017). In contrast to theories that interpret N400/P600 divergence as evidence for a dual-stream architecture, this model is offered as proof of concept for a single-stream account. The model instantiates the Retrieval-Integration (RI) theory (Brouwer et al., 2012), which holds that the N400 reflects lexical retrieval, while the P600 reflects integration of lexical content into the sentence meaning representation.

The success of this RI-based computational model in simulating divergent role reversal effects with a single-stream architecture is significant and worth examining in greater detail. Here we present a series of simulations aimed at replicating the Brouwer result and examining the factors driving its success. We find that the success of the model in capturing divergent sensitivity to reversal anomalies depends on potentially unrealistic properties of the training data, as well as presence of the test data in the training set. We discuss implications for interpretation of the model's results.

2 Brouwer model

Simulated experiment Brouwer et al. (2017) simulate Hoeks et al (2004), which showed N400/P600 divergence in role reversal anomalies. The Hoeks et al (2004) experiment contained four conditions: passive non-anomalous sentences, anomalous reversal sentences, and two types of anomalous "mismatch" sentences (see Table 1). All three anomalous conditions elicited P600 effects, while N400 effects were present only for the mismatch anomalies: the N400 showed insensitivity to the reversal.

Model The Brouwer model is a recurrent neural network, illustrated in Figure 1. It is trained in two phases, with error signal from target sentence representations at the final output layer. The integration and integration_output layers are first trained separately, with pre-trained word meaning vectors in place of input from the retrieval_output layer.[1] These integration layers are then frozen while the retrieval and retrieval_output layers are trained within the full model.[2] Inputs to the re-

[1]Brouwer uses 100-dimensional binary COALS vectors (Rohde et al., 2006).

[2]This process guides the retrieval_output layer toward the pre-trained word representations, which induces the retrieval

Test condition	Sentence (gloss of Dutch word order)	English translation
Passive (control)	the goal was by the player scored	"The goal was scored by the player."
Reversal	the goal has the player scored	"The goal scored the player."
Mismatch passive	the goal was by the player served	"The goal was served by the player."
Mismatch active	the goal has the player served	"The goal served the player."

Table 1: Illustration of simulation conditions, based on Hoeks et al. (2004). Sentences used in the original experiment and in all simulations reported here were in Dutch. Underlined final verbs indicate the critical word on which N400 and P600 measurements were taken.

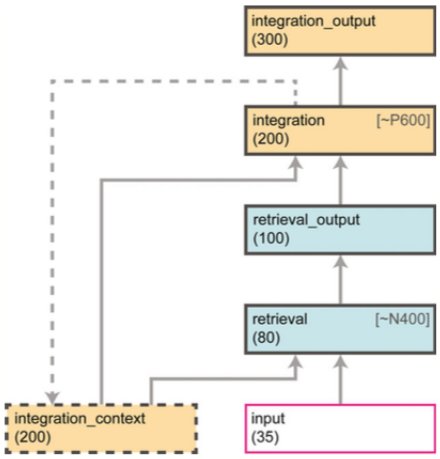

Figure 1: Retrieval-Integration model architecture (image credit: Brouwer et al. (2017))

trieval layer are localist (one-hot) word representations. Training labels for the integration_output layer are sentence representation vectors made up of three slots, to be filled by the word meaning vectors corresponding to the agent, action, and patient of the sentence.

As indicated in Figure 1, the simulated N400 amplitude is drawn from the retrieval layer, and the simulated P600 amplitude is drawn from the integration layer. These simulated amplitudes are taken to be the difference in activation (cosine distance) of the relevant layer from the previous timestep (word) to the current timestep (word). This means that in the model, the N400 is identified with the change, from w-1 to w, in a layer mapping (word identity + previous context) $\rightarrow$ word meaning. The P600 is identified with the change, from w-1 to w, in a layer mapping (word meaning + previous context) $\rightarrow$ sentence meaning.

Training data In each phase of training, Brouwer et al. train on data composed of two halves: eight-thousand "stereotypical" sentences, and eight-thousand of what we will call "all-combinations" sentences. The stereotypical half includes active and passive sentences reflecting each of ten non-anomalous agent-action-patient triplets, such as $\langle$*player scored goal*$\rangle$ and $\langle$*lawyer sued company*$\rangle$. These triplets are the basis of the passive control condition in the simulation. Each triplet is represented 800 times in this half of the data (400 active, 400 passive). By contrast, the all-combinations data represents every possible agent-patient-action combination from among the 20 nouns and 10 verbs ($20 \times 20 \times 10 = 4000$), each with an active and passive form (8000). The model thus trains on every possible combination of agents, patients and actions (e.g., $\langle$*goal sued lawyer*$\rangle$), but it sees the non-anomalous combinations with much greater frequency (at a ratio of 401:1). All test sentences are seen during training.

3 Replication

We implemented the RI model architecture, using the same two-phase training and number of training epochs. All simulations were run on Dutch data from Brouwer et al.'s Simulation 1.[3] Figures 2 and 3 show all simulation results.

The "Replication" setting of Figures 2 and 3 shows our results when training on the data described in Section 2. While the relative values of the mismatch conditions differed slightly from those reported in Brouwer's simulation (we find that these fluctuate slightly between runs), we have a consistent replication when it comes to the critical reversal condition: the N400 layer shows insensitivity to role reversal—the reversal condition patterns with the non-anomalous passive condition—while the P600 layer shows sensitivity to the reversal. Replicating Brouwer, our model also reaches 100% "comprehension" performance for output, where success (per Brouwer) means the output sentence vector has the highest cosine similarity with the target sentence vector when compared with all different training sentence vectors.

module to serve a retrieval-like function.

[3]We thank Harm Brouwer for generous discussion, clarification, and provision of original Dutch word vectors.

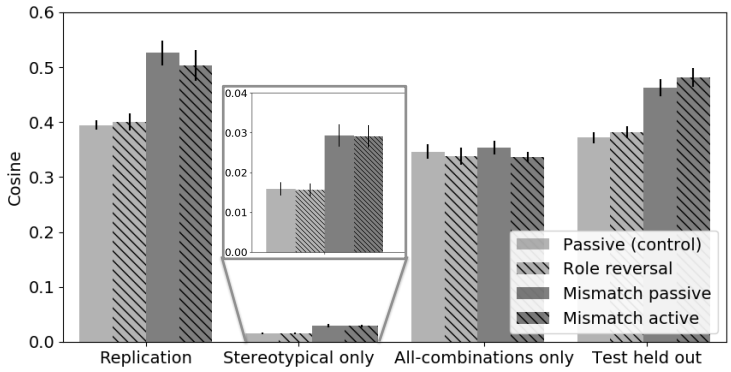

Figure 2: N400 simulation results across different training settings

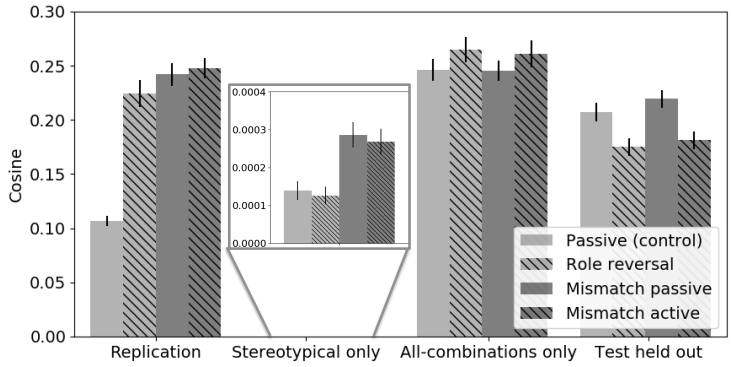

Figure 3: P600 simulation results across different training settings

4 Examining effects of training data

The data used to train this model is potentially problematic for two reasons. First, due to the all-combinations data, the training contains many implausible sentences that people are unlikely to have encountered prior to an experimental context. These implausible sentences include the anomalous meanings seen in the simulations, such as "the goal served the player", as well as a host of other implausible meanings, such as "the meal sang the painting". The authors justify the all-combinations data as teaching the model that any noun can serve either an agent or a patient role.

A second concern is that all simulation data is included in the training data, such that the model is not required to generalize. This means that the model's performance may be contingent on the distribution of test sentences within the training data—an unrealistic assumption for human cognition, given that humans frequently process sentences that they have not previously encountered.

Isolating effects of each half of the training set To isolate the contributions of each training data component, we ran separate simulations with each half. The "Stereotypical only" and "All-combinations only" settings in Figures 2 and 3 show simulation results when the model has been trained only on the corresponding half of the data.

The stereotypical-only simulation is of particular interest, as it allows us to test the model's performance with implausible sentences excluded from the training data. It is also the first setting in which the model is tested on partly unseen data (the non-anomalous passive condition is represented in the training data, but the three anomalous conditions are not). We find that comprehension accuracy drops from 100% to 25%, indicating that in this training setting the model does not generalize to comprehension of untrained sentences. Inspection verifies that the 25% of simulation sentences with correct output are those in the non-anomalous passive condition seen in training.

As for the N400/P600, we see that with stereotypical-only training, the model's N400 layer still captures the desired pattern, with attenuated N400 amplitude for reversals. Given that the target N400 pattern is what we would expect if a component were derived from a simple bag-of-words sentence model (c.f. Ettinger et al., 2016), it is not surprising that the simulation of this pattern persists. The P600 layer, by contrast, no longer shows the desired sensitivity to all three anomalies, instead resembling the N400 layer.

With all-combinations only, when implausible sentences are included in training and all sentences

are encountered with equal probability, simulation comprehension accuracy returns to 100%.[4] However, both the N400 and P600 layers now fail to simulate the desired effects, as all conditions pattern roughly together at both layers.

These results suggest that the model's simulation of the N400 pattern is driven not by the inclusion of implausible sentences, but rather by the relative training frequency of sentences with a given noun-noun-verb (NNV) combination, without regard to role (roughly a bag-of-words level of representation). This conclusion emerges from the fact that the N400 pattern is present when the model has seen only the NNV combinations shared by passive and reversal conditions (stereotypical-only), and when it has seen those NNV combinations with far greater frequency (full training set)—but not when all NNV combinations have been seen with uniform probability (all-combinations only).

By contrast, the results suggest that the model's simulation of the P600 effect requires the distribution provided by the full training dataset, as neither half independently produces the desired P600 pattern. Specifically, the P600 layer appears to reflect training frequency of particular sentences or role triplets—but only when trained with the combined stereotypical and all-combinations data.

Testing with held-out data The failure of generalization in the stereotypical-only setting may be due to the fact that the stereotypical sentences are distinguishable by very coarse features, and the model may thus not be forced to arrive at a sophisticated solution in training. To rectify this, we created a new dataset with sentences removed based only on the simulation data. The stereotypical-only half was converted to active sentences only, to avoid occurrences of the simulation's passive controls, and from the all-combinations half we removed all sentences reflecting role triples in the test data, both active and passive. The resulting training data has the original degree of skew toward stereotypical meanings, contains both actives and passives, and embodies the principle that any noun can be an agent or a patient.

In this setting, the simulation comprehension returns to 100%, indicating that when trained on fine-grained data, the model's comprehension does generalize to untrained sentences. However, we see in the "Test held out" setting in Figures 2 and 3 that the N400/P600 patterns now change, and in particular that the P600 layer, rather than reflecting the anomaly sensitivity, appears to reflect the training skew toward active sentences.[5]

Taken together, these results suggest that the model's success, particularly at the P600 layer, is driven by the relative frequency in training of the test sentences in each condition, rather than a more abstract representational factor that would allow for generalization of the effect to new sentences.

5 Conclusion and future work

We have provided a careful replication and deeper investigation of the single existing computational simulation of N400/P600 role reversal effects. We learn from our simulations that the success of this model is driven by the skew in the distribution of training sentences. Rather than reflecting abstraction, the model's P600 layer in particular patterns with relative frequencies of simulation sentences in training. This is problematic for interpreting the model's performance in terms of the RI theory, as the performance appears to be driven not (solely) by the RI-based architecture, but rather by properties of training which are neither included in the claims of the theory nor sufficiently cognitively plausible to be innocuous assumptions. In order to draw conclusions about the theories of interest, we need to ensure that the models make cognitively plausible assumptions and are robust to plausible variations not constrained by theoretical claims.

What it will take to build models that produce desired P600 effects on unseen sentences is an important matter for future work. We assume that such models will need to extract abstract properties from the learning process, in order to reflect different levels of anomaly in unseen sentences. This may be achievable as an emergent property of the learning, but it may also require representational capacities built into the model. This will be the subject of the next phase of our investigations.

References

Ina Bornkessel-Schlesewsky and Matthias Schlesewsky. 2008. An alternative perspective on seman-

[4] We did require more training epochs than other simulations in order to reach 100% comprehension.

[5] We find that occasionally the training loss spikes in the final epoch, an occurrence which is often accompanied by noisier N400 and P600 results. For most simulations we were able to avoid this, but due to erratic loss in this hold-out simulation, we used a diminishing learning rate as employed by Brouwer. Other simulations used a constant learning rate.

tic p600 effects in language comprehension. *Brain research reviews* 59(1):55–73.

Harm Brouwer, Matthew W. Crocker, Noortje J. Venhuizen, and John C. J. Hoeks. 2017. A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science* 41:1318–1352. https://doi.org/10.1111/cogs.12461.

Harm Brouwer, Hartmut Fitz, and John Hoeks. 2012. Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Research* 1446:127–143.

Wing-Yee Chow, Cybelle Smith, Ellen Lau, and Colin Phillips. 2016. A bag-of-arguments mechanism for initial verb predictions. *Language, Cognition and Neuroscience* 31(5):577–596.

Allyson Ettinger, Naomi H Feldman, Philip Resnik, and C Philips. 2016. Modeling n400 amplitude using vector space models of word representation. In *Proceedings of the 38th annual conference of the Cognitive Science Society*. Cognitive Science Society Austin, TX, pages 1445–1450.

John CJ Hoeks, Laurie A Stowe, and Gina Doedens. 2004. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research* 19(1):59–73.

Albert Kim and Lee Osterhout. 2005. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of memory and Language* 52(2):205–225.

Miriam Kos, Theo Vosse, Danielle Van Den Brink, and Peter Hagoort. 2010. About edible restaurants: conflicts between syntax and semantics as revealed by erps. *Frontiers in psychology* 1.

Gina R Kuperberg. 2007. Neural mechanisms of language comprehension: Challenges to syntax. *Brain research* 1146:23–49.

Douglas LT Rohde, Laura M Gonnerman, and David C Plaut. 2006. An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM* 8:627–633.

Marieke Van Herten, Herman HJ Kolk, and Dorothee J Chwilla. 2005. An erp study of p600 effects elicited by semantic anomalies. *Cognitive Brain Research* 22(2):241–255.