The Electrophysiology of Language Comprehension

Brouwer, Harm

CHAPTER 5

A Neurocomputational Model of the N400 and the P600 in Language Comprehension

Abstract | *One decade ago, researchers using event-related brain potential (ERP) measurements stumbled upon what looked like a Semantic Illusion in language comprehension: Semantically anomalous, but otherwise well-formed sentences did not affect the meaning-related N400 component, but instead increased the amplitude of the structure-related P600 component. This finding spawned five new models of language comprehension, all of which claim that instead of a single comprehension process, there are two or even more separate processing streams, one of which is not driven by structure, but by word meaning alone. Here, we will argue that there is a much simpler way to account for these data, and we will present evidence from neurocomputional simulations showing that our alternative explanation is able to predict all relevant ERP patterns found in the literature.*

5.1 Introduction

In electrophysiological research into language comprehension, there are two central brain responses. The first is the N400 component, a negative deflection of the ERP signal (Event-Related brain Potential) that peaks around 400 ms after stimulus onset, and that is sensitive to semantic anomalies such as 'He spread his warm bread with socks' (relative to *butter*; Kutas and Hillyard, 1980). The second is the P600 component, a positive deflection that can be maximal around 600 ms, and that was found in response to syntactic violations such as 'The spoilt child throw [...]' (relative to *throws*; Hagoort et al., 1993). The idea that semantic processing difficulty is reflected in the N400 component, and syntactic processing difficulty in the P600 component, has survived until this day. Ten years ago, however, findings emerged that presented a challenge to this mapping. Around 2003, more and more research groups discovered that certain types of syntactically sound, but semantically anomalous sentences failed to produce an N400-effect (but produced a P600-effect instead; e.g., Kolk et al., 2003; Kuperberg et al., 2003; Hoeks et al., 2004; Kim and Osterhout, 2005, among others). Hoeks et al. (2004), for instance, found that sentences such as 'De speer heeft de atleten geworpen' (lit: 'The javelin has the athletes thrown') produced an increase in P600 amplitude (relative to a non-anomalous control), but not in N400 amplitude. This was unexpected, because as javelins do not throw athletes, the word *thrown* should create semantic processing difficulty, and hence an increase in N400 amplitude. Equally unexpected was the finding of an effect on P600 amplitude in the absence of a syntactic anomaly. To account for these effects, increasingly complex models were proposed that incorporate multiple, potentially interacting processing streams (*Monitoring Theory*: Kolk et al., 2003; *Semantic Attraction*: Kim and Osterhout, 2005; *Continued Combinatory Analysis*: Kuperberg, 2007; the *extended Argument Dependency Model*: Bornkessel-Schlesewsky and Schlesewsky, 2008, and the *Processing Competition* account: Hagoort et al., 2009). What these models have in common is that they include a processing stream that is purely semantic, unconstrained by any structural

information (e.g., word order, agreement, case marking). This independent semantic analysis stream does not run into semantic processing problems on the word *thrown*, and hence does not produce an N400-effect, because the words *javelin*, *athletes*, and *thrown* fit together well semantically. Eventually, the processor does realize that something is wrong with the interpretation that was constructed, and the effort put into solving this problem (often structurally) is reflected in a P600-effect. Despite the attractiveness of these models, however, none of them seems capable of explaining the full range of relevant findings in the literature (see Brouwer et al., 2012, for a review).

5.1.1 A simpler perspective

In contrast to seeking an architectural explanation for the "Semantic Illusion" phenomenon (absence of an N400-effect in response to a semantic anomaly), Brouwer et al. (2012) argued for a functional reinterpretation of the ERP components involved. First of all, in line with others (e.g., Kutas and Federmeier, 2000; Lau et al., 2008; van Berkum, 2009), they suggested that the N400 component reflects *retrieval* of lexical semantic information, rather than compositional semantic processing or semantic integration. Retrieval of the information associated with a word is facilitated if that information is already (partly) activated by its prior context. This explains why the word *socks* engenders a much larger N400 in the context of 'He spread his warm bread with [...]' than the word *butter*: the lexical knowledge associated with *socks* is inconsistent with that already activated by its prior context (*socks* do not fit well with *spread* and *warm bread*), whereas the conceptual knowledge associated with *butter* is. It is important to note that under this account pre-activation can stem from the message representation that has been constructed so far (e.g., a breakfast scene), as well as from the preceding lexical items themselves (*spread* and *bread*).

In addition to this effect on N400 amplitude, presenting the word *socks* also leads to an increase in P600 amplitude, even though the sentence in

which it appears is grammatically correct (see Kutas and Hillyard, 1980, Figure C). The same is true for the "Semantic Illusion" sentences, where a P600-effect is elicited in a syntactically well-formed sentence. Brouwer et al. (2012) argued that these results indicate that the P600 is not a reflection of *syntactic* processing, but must represent some form of effortful *semantic* processing or integration. They put forward the hypothesis that the P600 component is actually a family of related components, all of which are late positivities that reflect the word-by-word construction, reorganization, or updating of a *mental representation of what is being communicated* (an MRC for short). Given this view on the P600 component, and the retrieval view on the N400 component, Brouwer et al. (2012) suggest that language is processed in biphasic N400/P600—Retrieval-Integration—cycles; every incoming word modulates the N400 component, reflecting the effort involved in activating its associated conceptual knowledge, as well as the P600 component, reflecting the effort needed to integrate this knowledge into an updated representation of what is being communicated. This parsimonious single stream account was shown to have the broadest empirical coverage of all the extant models (Brouwer et al., 2012), and indicated that the so-called "Semantic Illusion" in sentence processing, the effect that had led to so many new models, was just a simple instance of priming: if a preceding context (e.g., 'The javelin has the athletes [...]') pre-activates the lexical features of an incoming word (e.g, *thrown*), there is no N400-effect. On the other hand, when an incoming word is not consistent with the pre-activated lexical features (e.g., *summarized*), there is an N400-effect. Futhermore, P600 amplitude is expected to increase in every case where there is difficulty in making sense.

5.1.2 Mapping function to anatomy

Brouwer and Hoeks (2013) aligned these functional interpretations of the N400 and the P600 with neuroanatomy, and suggested that the processes of retrieval and semantic integration are mediated by specific brain areas. First of all, they propose that the left posterior Middle Temporal Gyrus

(lpMTG; BA 21) serves as a network *hub* that mediates lexical retrieval (∼N400). Such a hub (cf. Buckner et al., 2009), or epicenter (cf. Mesulam, 1990, 1998), is a brain region that serves to integrate or bind together information from various neighbouring areas (see also Damasio, 1989), and to broadcast this information across larger neuroanatomical networks. The lpMTG finds itself in the middle of brain areas that constitute long-term memory. Words (and also other meaningful stimuli) activate parts of that network, and the resulting information is collected through the lpMTG and shared with other brain networks for further (higher-level) processing. On the Retrieval-Integration account, this information is used to create a valid and coherent mental representation of what is communicated (an MRC). Brouwer and Hoeks (2013) suggested that the construction of such an MRC takes place in and around the left Inferior Frontal Gyrus (lIFG; BA 44/45/47). That is, the lIFG is a hub mediating MRC composition, and thus the most prominent source of the P600. Again, more areas may be involved in making sense, but the lIFG is the most central one, binding together the information from surrounding neural territory.

The information sharing between the two hubs (i.e., from the lpMTG to the lIFG and back) occurs via white matter tracts that structurally connect them. There are two major white matter pathways connecting the lpMTG and the lIFG, the dorsal pathway (dp) and the ventral pathway (vp), the precise functional roles of which are still poorly understood (see Brouwer and Hoeks, 2013, for discussion). Nonetheless, we can describe an approximate functional-anatomic Retrieval-Integration cycle of an incoming word (see Figure 5.1, top). First, a word reaches the lpMTG via either the auditory cortex (ac) or the visual cortex (vc), depending on the input modality. The lpMTG then retrieves the conceptual knowledge associated with this word from the association cortices and binds it together, a process that generates the N400 component. Next, the retrieved knowledge is shared with the lIFG, via one of the white matter pathways, where it is integrated with the prior context. This process is assumed to generate the P600 component. Finally, the new representation feeds back to the lpMTG causing

pre-activation of conceptual knowledge in memory.

In summary, compared to the five competing models, the Retrieval-Integration account is theoretically the most parsimonious, has the broadest empirical coverage, and seems to fit well with what is presently known about the neuroanatomy of language. However, what it has in common with the other models, is that it is still a conceptual *box-and-arrow* model. This means that the predicted outcome of the model in any specific case is at best *qualitative*, and may be affected by implicit biases and other subjective factors; that is, researchers may for instance disagree on the amount of contextual and lexical priming in any specific sentence, and their predictions on the presence or absence of an N400-effect may vary. The only way to overcome this problem is to implement the model computationally, which means giving it a formally precise description of the mechanisms that are supposed to underlie it, and then running this computational model to generate *quantitative* predictions. The predictions in terms of N400 amplitude and P600 amplitude, can then be compared to the actual results of empirical studies.

5.1.3 A neurocomputional model

We present a neurocomputational model that predicts the amplitude of the N400 and the P600 component at every word of a sentence. This model directly instantiates the functional-anatomic mapping of the Retrieval-Integration account that we described above. Following our hypothesis on the neuroanatomical organization of the language processor, the computational model consists of two connected but relatively independent subsystems: A system for retrieval (∼lpMTG) and a system for integration (∼lIFG). The retrieval system is trained to map words onto their lexical-semantic representations (extracted from a corpus using the Correlated Occurrence Analogue to Lexical Semantics, COALS; Rohde et al., 2009). The integration system, in turn, is trained to map these lexical-semantic representations onto an approximation of the 'meaning' of a sentence, a thematic role assignment in terms of the *agent*, *action*, and *patient* (i.e., *who-*

did-*what*-to-*whom/what*). Importantly, the mapping of words onto their lexical-semantic representations in the retrieval system (∼lpMTG), can be facilitated by lexical and higher level cues that are present in the unfolding representation in the integration system (∼lIFG).

In what follows, we will first present the model conceptually, and show that it can produce the same patterns of ERPs as found in an actual empirical study. In the 'Methods' section, we will provide the full technical details of the model.

5.2 Results

5.2.1 The neurocomputational model

Our neurocomputational model consists of five layers of artificial neurons, one corresponding to the auditory/visual cortex (ac/vc), two for the left posterior MTG (lpMTG and lpMTG_output), and two for the left IFG (lIFG and lIFG_output) (see Figure 5.1, bottom). Activation flows forward from the ac/vc layer all the way to the lIFG_output layer, and from the lIFG layer both back into itself (in order to update the MRC at the previous time-step) and to the lpMTG layer (providing a context for retrieval). The model is taught that any noun phrase can theoretically be an *agent* or a *patient*, but that there are certain stereotypical combinations of *agents*, *patients*, and *actions* (∼minimal world knowledge, see also Mayberry et al., 2009).

5.2.2 Linking hypotheses

In our view, N400 amplitude is a measure of 'unpreparedness'. If no features relevant to an incoming word are pre-activated, N400 amplitude will be maximal; if the lexical-semantic features of an incoming word are consistent with those pre-activated in memory, N400 amplitude will be reduced. Hence, N400 amplitude is a measure of how much the activation pattern in memory changes due to the processing of an incoming word. As such, we compute the correlates of N400 amplitude at the lpMTG layer,

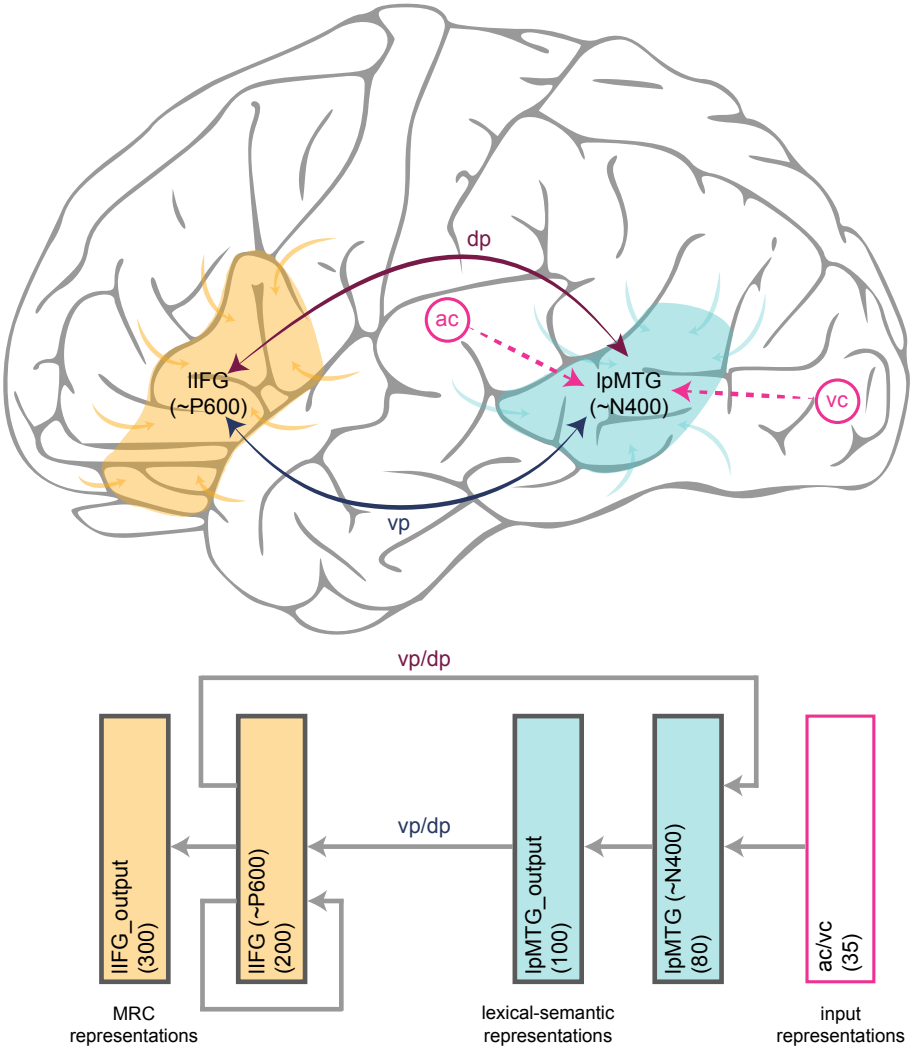

Figure 5.1 | *Schematic illustration of the neurocomputational model (bottom), and its relation to the functional-anatomic mapping of the core language network (top). An incoming word reaches the lpMTG via either the auditory cortex or visual cortex (ac/vc). The lpMTG then retrieves the conceptual knowledge associated with this word from the association cortices (lpMTG → lpMTG_output), thereby generating the N400. Next, this retrieved meaning is sent to the lIFG (lpMTG_output → lIFG), where it is integrated with its prior context (lIFG → lIFG), into an updated MRC (lIFG → lIFG_output). The updated MRC in the lIFG subsequently provides a context for the retrieval of the conceptual knowledge assocatied with the next word (lIFG → lpMTG).*

Item	Condition	Effect
De speer werd door de atleten geworpen *The javelin was by the athletes thrown*	Control (Passive)	—
De speer heeft de atleten geworpen *The javelin has the athletes thrown*	Reversal (Active)	P600
De speer werd door de atleten opgesomd *The javelin was by the athletes summarized*	Mismatch (Passive)	N400/P600
De speer heeft de atleten opgesomd *The javelin has the athletes summarized*	Mismatch (Active)	N400/P600

Table 5.1 | *Materials used in the original ERP experiment, as well as the effects that were observed for each condition.*

where the activation of lexical-semantic features takes place (∼memory retrieval), as the degree to which the pattern of activity induced by the current word, and that induced by the previous word are *different* (see 'Methods' section for mathematical details).

P600 amplitude, in turn, reflects the difficulty of establishing coherence. The more the current interpretation (the current MRC) needs to be reorganized or augmented in order to become coherent, the higher P600 amplitude. Hence, P600 amplitude is effectively a measure of how much the representation of the unfolding state of affairs changes due to the integration of an incoming word. As such, we compute the correlates of P600 amplitude as the difference between the previous and the current state of affairs at the lIFG layer, where the (re)construction of an MRC—in terms of thematic-role assignment—takes place (see also Crocker et al., 2010).

5.2.3 Modeling Event-Related Potentials

The patterns of N400 and P600 elicitation reported in the literature (see Kutas et al., 2006, for an overview) suggest that there are two main processing outcomes. The most common is the biphasic pattern where an N400-effect co-occurs with a P600-effect. When context does not prepare for a specific upcoming word, retrieval is more difficult (hence an N400-effect), and often the retrieved information is more or less unexpected, leading to

more effortful MRC construction (P600-effect). Less frequent are cases that elicit only a P600-effect, such as in "Semantic Illusion" sentences. There, the retrieval of a word is facilitated by context or preceding lexical items (hence no N400-effect), but integration is problematic (P600-effect). Theoretically, there is also a third option: retrieval is more involved, but integration is as easy as in a control sentence (hence only an N400-effect, and no P600-effect). However, isolated N400-effects are scarce, and their presence may be a baseline artefact (Brouwer and Hoeks, 2013; Hoeks and Brouwer, 2014).

Our computational model should thus be able to simulate the two most common findings: biphasic N400/P600-effects as well as isolated P600-effects. Both types of outcomes are present in the study by Hoeks et al. (2004), which we will take as reference point. This study compared semantically anomalous Dutch sentences like 'De speer *heeft* de atleten geworpen' (lit: 'The javelin *has* the athletes thrown') to normal controls like 'De speer *werd door* de atleten geworpen' (lit: 'The javelin *was by* the athletes thrown'). This comparison produced a P600-effect on the final verb *thrown*, but no N400-effect. Two other semantically anomalous conditions were also compared to the same control, which both gave rise to a biphasic N400/P600-effect on the final word: 'De speer *heeft* de atleten opgesomd' (lit: 'The javelin *has* the athletes summarized') and 'De speer *werd door* de atleten opgesomd' (lit: 'The javelin *was by* the athletes summarized'). Table 5.1 provides an overview of these materials and findings, and Figure 5.2 shows the results at the Pz electrode.

5.2.4 Simulation experiments

We modeled the results of the Hoeks et al. (2004) study in two simulation experiments. To assure that any effects that we find are not an artifact of the materials that we used, we used different sets of lexical items, and hence different sets of lexical-semantic representations, in the two simulation experiments.

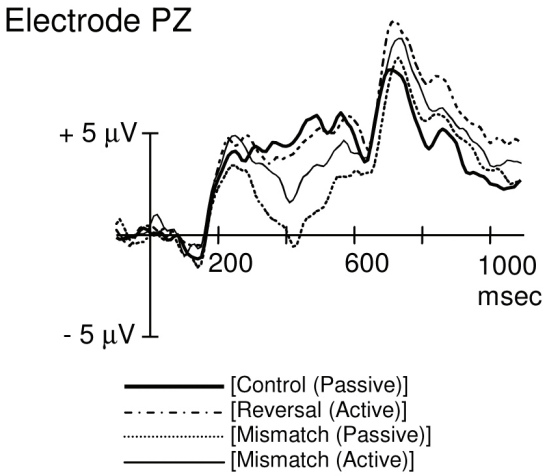

Figure 5.2 | *Results of the original ERP experiment on the Pz electrode. Positive is plotted upwards. Note that the reported ERP effects are based on multiple electrodes, and that this single electrode only serves as a representative example.*

5.2.4.1 N400 results

The computational model should be able to produce two very different findings in the N400 domain: an absence of an N400-effect in the comparison of "Semantic Illusion" sentences ('The javelin *has* the athletes thrown') with normal controls, and presence of an N400-effect in the mismatch anomaly conditions ('The javelin *has/was by* the athletes summarized'). Figure 5.3 shows the N400 results of the two simulation experiments (bottom graphs) in comparison to those of the original ERP experiment (top). Statistical evaluation using Repeated Measures ANOVA (with Condition as four-level within-items factor and Huynh-Feldt correction where necessary) showed a close replication of the Hoeks et al. (2004) findings. The main effect of Condition was significant in each of the simulation experiments (Exp 1: $F(3,27)=35.2$; $p<.001$; Exp 2: $F(3,27)=15.3$; $p<.001$), and pairwise comparisons (Bonferroni corrected) showed that 1) N400 amplitude

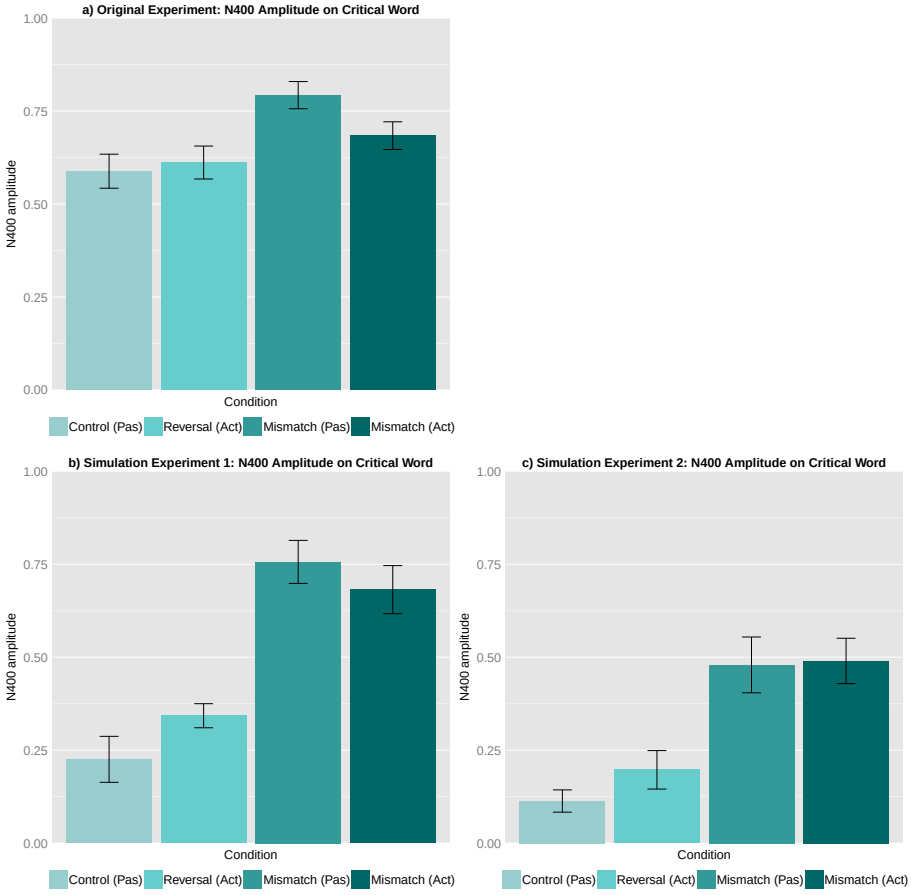

Figure 5.3 | *N400 results of the simulations in comparison to the results of the original experiment. Panel (a) shows the N400 amplitudes as measured in the original experiment (at the Pz electrode). Panel (b) shows the N400 correlates measured in simulation experiment 1, and panel (c) those measured in simulation experiment 2. All scores are transformed to a common scale. Error bars show standard errors of these transformed scores.*

did not differ between "Semantic Illusion" sentences and controls (Exp 1: p=.47; Exp 2: p=.91), and 2) there was a significant N400-effect for the two other anomalous conditions (Exp 1: p-values<.001; Exp 2: p-values<.005).

5.2.4.2 P600 results

For the P600 component, the main result that we wanted to replicate is an increase in amplitude for all anomalous conditions relative to control. Figure 5.4 shows the P600 results of the two simulation experiments (bottom graphs) in comparison to those of the original ERP experiment (top left). Again, we clearly replicated the main result; in both simulations, P600 amplitude was significantly higher for all anomalous sentences (including the "Semantic Illusion" sentences) compared to controls (Main effect of Condition: Exp 1: $F(3,27)=118.9$; $p<.001$; Exp 2: $F(3,27)=57.1$; $p<.001$); pairwise comparison showed that there was a significant P600-effect for all three anomalous conditions compared to control (Exp 1: p-values<.001; Exp 2: p-values<.001).

One slight difference between the actual and simulated results lies in the ordering of the three implausible conditions. The computational model predicts P600 amplitude to be largest for both *mismatch* conditions, whereas in the original ERP experiment, it is largest for the "Semantic Illusion" condition. A possible reason for this difference is that in the ERP experiment the pattern of P600-effects is affected by an artefact caused by *component overlap* (see Hagoort, 2003; Brouwer and Hoeks, 2013, for discussion), where the size of the preceding N400 affects the size of the P600. Figure 5.4 (top right) shows that if we correct for component overlap (by subtracting N400 amplitude from P600 amplitude in each condition), the pattern of results within the three anomalous conditions comes in line with the results of the simulations (mismatch conditions larger than "Semantic Illusion" sentences).

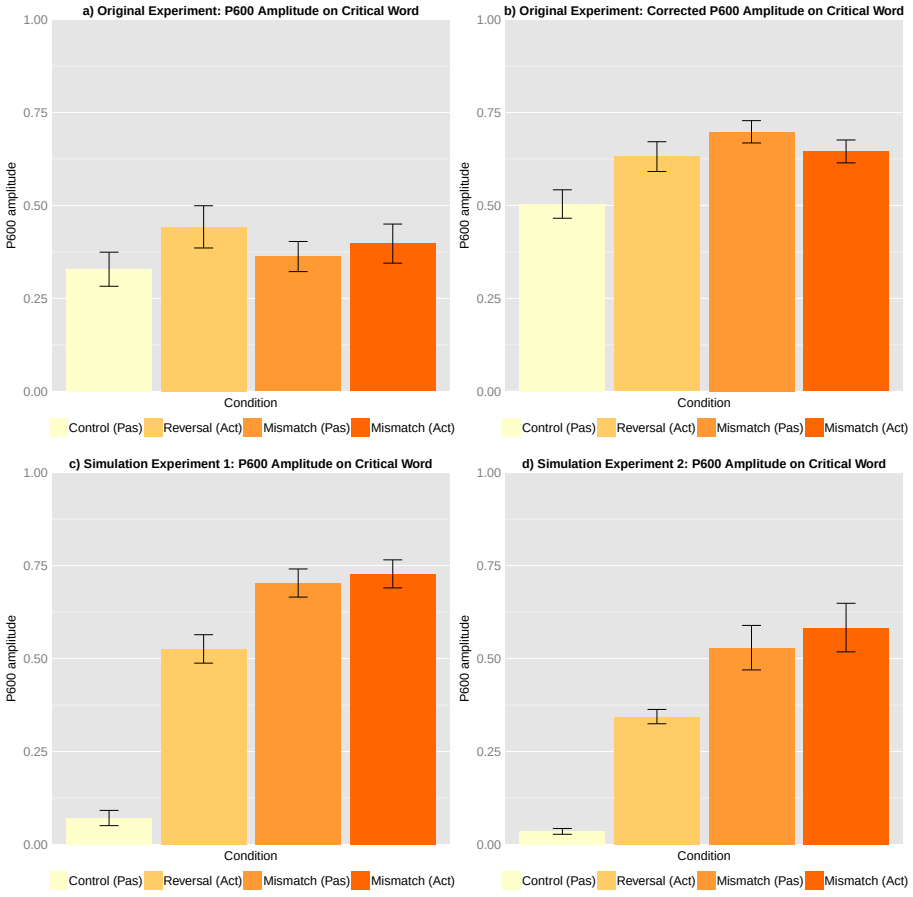

Figure 5.4 | *P600 results of the simulations in comparison to the results of the original experiment. Panel (a) shows the P600 amplitudes as measured in the original experiment (at the Pz electrode). Panel (b) shows the P600 amplitudes corrected for overlap with the N400 component (also at Pz). Panel (c) shows the P600 correlates measured in simulation experiment 1, and panel (d) those measured in simulation experiment 2. All scores are transformed to a common scale. Error bars show standard errors of these transformed scores.*

5.3 Discussion

We presented a neurocomputational model of language processing that directly instantiates a simple functional-anatomic mapping of the core language network. In our model, the N400 component and the P600 component reflect distinct processes that take place in distinct cortical regions; the N400 component reflects lexical retrieval processes mediated by the lpMTG, and the P600 component reflects integrative processes mediated by the lIFG. The computational model was trained to honor this division of labor by means of a two-stage training procedure (see 'Methods' section). The model was able to simulate the two most important patterns of ERPs as reported in the literature (P600-effects and biphasic N400/P600-effects), thereby providing a 'proof of concept' of our Retrieval-Integration view (cf. Brouwer and Hoeks, 2013).

5.3.1 Implications for other models

Our simulations have several implications for previously proposed neurocognitive processing models (Kolk et al., 2003; Kim and Osterhout, 2005; Kuperberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008; Hagoort et al., 2009). First of all, our simulations confirm that there is no need for an independent semantic analysis stream to explain "Semantic Illusions" in sentence processing, or "Semantic P600"-effects (cf. Brouwer et al., 2012). Our model simply proposes that there is a continuous process of making sense that takes place in the lIFG and that generates the P600, the amplitude of which is proportional to the amount of effort needed to (re-)construct an MRC. Each time a word (or other meaningful stimulus) comes in, this triggers a memory search for information associated with that stimulus, a search mediated by the lpMTG, which generates the N400, the amplitude of which reflects the amount of effort needed to retrieve this meaningful information. If the meaning of incoming stimulus is primed, retrieval will be easy, and N400 amplitude will be small (less negative). This retrieved information is then used by the lIFG to up-

date the mental representation of what is communicated. Our model is both architecturally more parsimonious than previously proposed models, while having broader empirical coverage (see Brouwer et al., 2012, Table 2). Secondly, previously proposed models have been developed as 'box-and-arrow' models, limiting them to *qualitative* predictions about the presence or absence of ERP effects. Instantiating our functional-anatomic mapping as a neurocomputational model, by contrast, adds a *quantitative* dimension to the predictions that it makes. A clear example here, is that whereas all other models simply predict absence of an N400-effect for the reversal condition (Brouwer et al., 2012, Table 2), we predict a slight, albeit non-significant increase in N400 amplitude relative to control. Critically, this slight increase is also present in the original ERP data. Hence, in addition to being more parsimonious, our model also makes more fine-grained predictions.

5.3.2 Future directions

The present computational model represents just a first step towards a full neurocomputational model of language processing. It will be necessary to expand our simulations in several respects. For one, we want to model other types of anomalous sentences. In addition, we want to model P600-effects arising from other processing phenomena, such as syntactic violations (see Gouvea et al., 2010, for an overview), and pragmatic violations (see Hoeks et al., 2013; Hoeks and Brouwer, 2014). Also, we want to include other ERP components into our model, such the Early Left Anterior Negativity (ELAN), which is elicited by word category violations, and the Left Anterior Negativitiy (LAN), which appears to be related to the processing of morphosyntactic marking (see Friederici, 2011). A completely different direction for future exploration is to expand the number of ERP component features covered by our model. For now we have focused on component amplitude (and implicitly polarity), but a component is also defined by its *latency*, *duration*, and *scalp distribution*. It remains to be seen how these features can be incorporated into our model.

5.4 Methods

5.4.1 Model Architecture and Activation Flow

The model is essentially an extended Simple Recurrent Network (SRN) (Elman, 1990) consisting of five layers: ac/vc (35 units), lpMTG (80), lMTG_output (100), lIFG (200), and lIFG_output (300) (see Figure 5.1). Input patterns are clamped to the ac/vc layer, and activation propagates feed forward to the lIFG_output layer following the trajectory: ac/vc $\rightarrow$ lpMTG $\rightarrow$ lpMTG_output $\rightarrow$ lIFG $\rightarrow$ lIFG_output. In addition, at a given processing timestep t, the lIFG and the lpMTG also receive input from the activation pattern of the lIFG at timestep $t-1$ through an additional 200-unit context layer that receives a copy of the activation pattern of the lIFG prior to feed forward propagation. At timestep $t = 0$, the activation value of each unit in this context layer was set to 0.5. All layers except this context layer and the ac/vc layer also receive input from a bias unit, the value of which is always 1. Unit activation y_j of a unit j is determined by a sigmoid activation function:

$$y_j = \frac{1}{1 + e^{-x_j}} \tag{5.1}$$

where x_j is the net input to unit j:

$$x_j = \sum_i y_i w_{ij} \tag{5.2}$$

which is determined by the activation level y_i of each unit i that propagates to unit j, and the weight w_{ij} on the connection between these units.

The model was trained in two stages (see below for detailed information on the training procedure). In the first stage, the ac/vc layer and the lpMTG layer were excluded from the network, and the model was trained to map sequences of *lexical-semantic representations* (see below for details on the representations used), clamped to the lpMTG_output layer, to *thematic-role assignment representations* in the lIFG_output layer. In the second stage, the ac/vc layer and the lpMTG layer were included, and the

model was trained to map sequences of *acoustic/orthographic representations* clamped to the ac/vc layer to thematic-role assignment represention in the lIFG_output layer. Critically, all the weights that were present in the network during the first training stage were frozen, forcing the network to first map the acoustic/orthographic representations to lexical-semantic representations in the lpMTG_output layer, before mapping these to thematic-role assignment representations in the lIFG_output layer.

5.4.2 Training and Testing Patterns

For both experiments, we created two sets of training items (one for both training stages), and a set of test items. The training items consist of Dutch active and passive sentences with the following template structure:

Active sentences:

De	[AGENT]	heeft	het/de	[PATIENT]	[ACTION]
The	[AGENT]	has	$\text{the}_{(+/-\text{NEUTER})}$	[PATIENT]	[ACTION]

Passive sentences:

De	[PATIENT]	werd	door	het/de	[AGENT]	[ACTION]
The	[PATIENT]	was	by	$\text{the}_{(+/-\text{NEUTER})}$	[AGENT]	[ACTION]

From these templates, training sentences were generated by filling in the *agent*, *patient*, and *action* slots using the noun phrases (agent and patient) and verbs (action) listed in Table 5.2. Half of the training sentences were constructed by permuting each of the twenty noun phrases (agents plus patients) with each verb ($2 \times 20 \times 20 \times 10 = 8000$ items). The other half consisted only of sentences with stereotypical agent-patient-action combinations, which are listed in the rows of Table 5.2. Hence, there are a total of 16000 (2×8000) training items, in which each verb appears 1600 times, 802 times ($\approx 50\%$) of which in a stereotypical agent-patient-action construction. As the model processes sentences word-by-word, each training item consists of a sequence of either 6 (active sentences) or 7 (passive sentences) pairs of input and target patterns. The input patterns consist of either lexical-semantic representations (stage one), or acoustic/orthographic

representations (stage two). The target patterns, in turn, are always the desired thematic-role assignment representation for a sentence.

The test sets consist of 40 sentences that are evenly divided over four conditions. There are 10 passive stereotypical agent-patient-action sentences (controls), 10 active role-reversed sentences (reversals), which were constructed by swapping the stereotypical agents and patients, and 10 active as well as 10 passive mismatch sentences (active and passive mismatches), in which the stereotypical action verb is replaced by the mismatch verb listed in Table 5.2.

5.4.3 Representations

5.4.3.1 Acoustic/orthographic vectors

The acoustic/orthographic input representations are 35-unit localist vectors, in which each unit corresponds to a single word (20 nouns + 10 verbs + 2 auxiliary verbs + 2 determiners + 1 preposition = 35 words).

5.4.3.2 Lexical-semantic vectors

The lexical-semantic representations are 100-unit binary representations, which were derived from a large corpus of Dutch newspaper texts using the Correlated Occurrence Analogue to Lexical Semantics (COALS; Rohde et al., 2009). In all of the steps described below, we precisely follow the procedure laid out in Rohde et al. to derive these representations (or COALS vectors).

We first derived a co-occurrence matrix using a 4-word ramped window, meaning that a word a co-occurs with b if a occurs within 4 words to the left or right of b, and that this co-occurrence is weighted by the proximity of a to b on a scale of 4 (direct neighbor) to 1 (separated by three words). We then pruned all but the 14.000 columns representing the most frequent words, so that the rows of the matrix then represented 14K-dimensional word feature vectors. Next, the weighted frequency of each co-occurrence

$w_{a,b}$ of words a and b was normalized by converting it to a pairwise correlation:

$$w'_{a,b} = \frac{T \cdot w_{a,b} - \sum_j w_{a,j} \cdot \sum_i w_{i,b}}{(\sum_j w_{a,j} \cdot (T - \sum_j w_{a,j}) \cdot \sum_i w_{i,b} \cdot (T - \sum_i w_{i,b}))^{\frac{1}{2}}} \tag{5.3}$$

where i is a row index, j is a column index, and:

$$T = \sum_i \sum_j w_{i,j} \tag{5.4}$$

In the resulting matrix, we replaced each negative correlation with 0, and each positive correlation with its square root:

$$norm(w'_{a,b}) = \begin{cases} 0 & \text{if } w'_{a,b} < 0 \\ \sqrt{w'_{a,b}} & \text{otherwise} \end{cases} \tag{5.5}$$

To obtain the 100-dimensional feature vectors that we used in our simulations, we pruned all but the 15.000 rows representing the most frequent words, and then reduced the dimensionality of the normalized feature vectors for these words by computing the Singular Value Decomposition of the co-occurrence matrix $X_{15000\times14000}$. Here we considered only the first 100 singular values and vectors, such that we obtain matrix $\hat{X}$ that is the best rank-100 approximation to X in terms of sum squared error:

$$\hat{X}_{15000\times14000} = \hat{U}_{15000\times100}\hat{S}_{100\times100}\hat{V}^T_{100\times14000} \tag{5.6}$$

A 100-unit feature vector V_c for a word c is then defined as:

$$V_c = X_c\hat{V}\hat{S}^{-1} \tag{5.7}$$

which can be converted to a binary vector by setting its negative components to 0, and its positive components to 1.

5.4.3.3 Thematic-role assignment vectors

The thematic-role assignment representations are 300-unit vectors, which are divided into three 100-unit slots. These three slots respectively represent the lexical-semantic representations of the elements that will be *agent*, *action*, and *patient* (cf. Mayberry et al., 2009).

5.4.4 Training and Testing

In both stages, the model was trained using bounded gradient descent (Rohde, 2002), a modification of the standard backpropagation algorithm (Rumelhart et al., 1986a).

For each input-target pair c, we minimized the sum squared error E_c between the desired activity d_j and the observed activity y_j for each unit j in the lIFG_output layer:

$$E_c = \frac{1}{2}\sum_j (y_j - d_j)^2 \tag{5.8}$$

Error was reduced by adjusting each weight w_{ij} in the model on the basis of a delta that is proportional to the gradient of that weight, and depends on its previous delta:

$$\Delta w_{ij}(t) = -\varepsilon\rho\frac{\partial E}{\partial w_{ij}} + \alpha\Delta w_{ij}(t-1) \tag{5.9}$$

where ε is the network's *learning rate*, ρ a *scaling factor* that depends on the length of the entire gradient:

$$\rho = \begin{cases} \frac{1}{||\partial E/\partial w||} & \text{if } ||\partial E/\partial w|| > 1 \\ 1 & \text{otherwise} \end{cases} \tag{5.10}$$

and α a momentum coefficient, controlling the fraction of the previous weight delta to be added.

The gradient $\frac{\partial E}{\partial w_{ij}}$ of a weight w_{ij}, in turn, is estimated as the product of the *error signal* δ_j of a unit j, and the activation value y_i of a unit i that signals to unit j:

$$\frac{\partial E}{\partial w_{ij}} = \delta_j y_i \tag{5.11}$$

The error signal δ_j for an output unit j is defined as:

$$\delta_j = (y_j - d_j)(y_j(1 - y_j) + 0.1) \tag{5.12}$$

where the constant 0.1 is a flat spot correction constant, preventing the derivative $y_j(1 - y_j)$ of the sigmoid activation function to approach zero when y_j is near 0 or 1 (cf. Fahlman, 1988). The error signal δ_j for a hidden unit j, in turn, is defined as:

$$\delta_j = (y_j(1 - y_j) + 0.1) \sum_k \delta_k w_{jk} \tag{5.13}$$

where all units k are units that receive signals from unit j.

We trained the model for 7000 epochs, in each of which we accumulated gradients over 100 input-output pairs before updating the weights. Training items were presented in a permuted order, such that by the end of training, the model has seen each item at least 43 times ($7000/(16000/100) = 43.75$). After all of the 16000 items were presented once, the training order was permuted again. Weights were initially randomized within a range of $(-0.25, +0.25)$, and were updated using a learning rate ε of 0.2, which was scaled down to 0.11 with a factor of 0.95 after each 700 epochs (that is, after each 10% interval of the total epochs; $0.2 \times 0.95^{10} \approx 0.11$). The momentum coefficient α was set to a constant of 0.9. Finally, we used a *zero error radius* of 0.1, such that no error was backpropagated if $(y_j - d_j) < 0.1$. The training procedure was identical for stage one and two.

After training, we evaluated the comprehension performance of the model using a output-target similarity matrix. For each item, we computed the cosine similarity between the output vector for that item, and each of the 16000 different target vectors. The cosine similarity between two vectors is defined as:

$$cos(x, y) = \frac{\sum_i x_i \times y_i}{\sqrt{(\sum_i x_i^2)} \times \sqrt{(\sum_i y_i^2)}} \tag{5.14}$$

The output vector for an item was considered correct if it was more similar to its corresponding target vector than to the target vector of any other item. We computed comprehension performance after stage one on the training set (at the lpMTG_output layer) and after stage two on both the training and the test set (at the lIFG_output layer). In all cases, comprehension performance was perfect (100% correct) for both experiments.

5.4.5 Computing ERP correlates

The ERP correlates for the N400 component were computed as the cosine dissimilarity between the lpMTG vector at time-step t and the lpMTG vector at time-step $t-1$:

$$\text{N400} = 1 - cos(lpMTG_t, lpMTG_{t-1}) \tag{5.15}$$

The P600 correlates were computed in a similar fashion at the lIFG layer:

$$\text{P600} = 1 - cos(lIFG_t, lIFG_{t-1}) \tag{5.16}$$

Exp.	Agent	Patient	NEUTER	Action	Mismatch
1	voetballer *soccer player*	doelpunt *goal*	+	gescoord *scored*	gediend *served*
1	militair *soldier*	land *country*	+	gediend *served*	gescoord *scored*
1	kok *cook*	maaltijd *meal*	-	bereid *prepared*	gezongen *sung*
1	zanger *singer*	lied *song*	+	gezongen *sung*	bereid *prepared*
1	advocaat *lawyer*	bedrijf *company*	+	aangeklaagd *sued*	gelopen *ran*
1	atleet *athlete*	marathon *marathon*	-	gelopen *ran*	aangeklaagd *sued*
1	politicus *politician*	debat *debate*	+	gevoerd *engaged*	uitgegeven *published*
1	uitgever *publisher*	roman *novel*	-	uitgegeven *published*	gevoerd *engaged*
1	arts *doctor*	diagnose *diagnosis*	-	gesteld *made*	geschilderd *painted*
1	schilder *painter*	schilderij *painting*	+	geschilderd *painted*	gesteld *made*
Exp.	**Agent**	**Patient**	NEUTER	**Action**	**Mismatch**
2	rechercheur *detective*	moord *murder case*	-	opgelost *solved*	verhoogd *raised*
2	werkgever *employer*	salaris *salary*	+	verhoogd *raised*	opgelost *solved*
2	dief *thief*	museum *museum*	+	beroofd *robbed*	getrokken *pulled*
2	tandarts *dentist*	tand *tooth*	-	getrokken *pulled*	beroofd *robbed*
2	schipper *sailor*	schip *ship*	+	aangelegd *berthed*	geregisseerd *directed*
2	regisseur *director*	film *movie*	-	geregiseerd *directed*	aangelegd *berthed*
2	piloot *pilot*	vliegtuig *airplane*	+	bestuurd *steered*	afgelegd *taken*
2	student *student*	tentamen *examen*	+	afgelegd *taken*	bestuurd *steered*
2	verzekeraar *insurer*	verzekering *insurance*	-	uitgekeerd *paid*	gereden *rode*
2	wielrenner *cyclist*	etappe *stage*	+	gereden *rode*	uitgekeerd *paid*

Table 5.2 | *Overview of the materials used in the simulation experiments. See text for details.*