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

Harm Brouwer[1,2], Matthew W. Crocker[1], Noortje J. Venhuizen[1] and John C. J. Hoeks[2]

[1]Department of Computational Linguistics and Phonetics (Psycholinguistics), Saarland University, Building C7.1, 66123 Saarbrücken, Germany.

[2]Center for Language and Cognition Groningen, University of Groningen, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands.

Correspondence should be addressed to: Harm Brouwer (brouwer@coli.uni-saarland.de)

Keywords: Language comprehension; Event-Related Potentials (ERPs); N400; P600; Retrieval-Integration account; Computational modeling; Neural Networks

Abstract

Ten years ago, researchers using event-related brain potentials (ERPs) to study language comprehension were puzzled by what looked like a *Semantic Illusion*: Semantically anomalous, but structurally well-formed sentences did not affect the N400 component—traditionally taken to reflect semantic integration—but instead produced a P600-effect, which is generally linked to syntactic processing. This finding led to a considerable amount of debate, and a number of complex processing models have been proposed as an explanation. What these models have in common is that they postulate two or more separate processing streams, in order to reconcile the Semantic Illusion and other semantically induced P600-effects with the traditional interpretations of the N400 and the P600. Recently, however, these multi-stream models have been called into question, and a simpler single-stream model has been proposed. According to this alternative model, the N400 component reflects the retrieval of word meaning from semantic memory, and the P600 component indexes the integration of this meaning into the unfolding utterance interpretation. In the present paper, we provide support for this 'Retrieval–Integration' account by instantiating it as a neurocomputational model. This neurocomputational model is the first to successfully simulate N400 and P600 amplitude in language comprehension, and simulations with this model provide a proof of concept of the single-stream Retrieval–Integration account of semantically-induced patterns of N400 and P600 modulations.

1 Introduction

In electrophysiological research into language comprehension, two brain responses take center stage. The first is the N400 component, a negative deflection of the Event-Related brain Potential (ERP) signal. It peaks around 400 ms after stimulus onset, and 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 generally reaches maximum around 600 ms. This component can be found in response to syntactic violations (such as 'The spoilt child throw . . . ', relative to 'throws' Hagoort et al., 1993). The dissociative sensitivity of the N400 to semantics and the P600 to syntax has led to the tenet that the N400 indexes processes of semantic integration, whereas the P600 indexes processes of a syntactic nature (see Kutas et al., 2006, for an overview). However, this mapping, which is at the core of many neurocognitive models of language processing, has been challenged by a set of findings that started to accumulate over the last decade. A number of studies revealed that certain types of syntactically sound, but semantically anomalous sentences failed to elicit the expected 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). For instance, Dutch sentences such as 'De speer heeft de atleten geworpen' (lit: 'The javelin has the athletes thrown', meaning "the javelin threw the athletes") produced an increase in P600 amplitude, but not in N400 amplitude, relative to a non-anomalous control 'De speer werd door de atleten geworpen' (lit: 'The javelin was by the athletes thrown', meaning 'the javelin was thrown by the athletes'; Hoeks et al., 2004). Provided the established views on the N400 and the P600, findings such as these came as a surprise. That is, as javelins cannot throw athletes, the word *thrown* should create semantic processing difficulty, and hence an increase in N400 amplitude. Also, as there is no need for syntactic reanalysis, there should be no increase in P600 amplitude. In search of an explanation for these 'Semantic Illusion'- or

'Semantic P600'-effects[1] that maintains the established views on the N400 and the P600, the literature has seen a shift towards so-called multi-stream models of language processing, in which a structure-insensitive semantic analysis stream operates in parallel to (and potentially interacts with) a more structure-driven algorithmic processing stream (see Kuperberg, 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012, for reviews).

In the present paper, we put forward an alternative, single-stream account of these 'Semantic P600'-effects, for which we provide explicit computational support. That is, we present a formally precise neurocomputational 'neural network' model that instantiates the recent Retrieval–Integration account of the N400 and the P600 in language comprehension (Brouwer et al., 2012; Brouwer and Hoeks, 2013). We show that our neurocomputational simulations produce relevant patterns of N400 and P600 effects reflecting semantic processing, including the 'Semantic P600'-effect. We argue that these results provide a proof of concept of the Retrieval–Integration account.

2 The Retrieval–Integration account

The finding that certain types of semantically anomalous, but syntactically sound sentences, do not produce an increase in N400 amplitude, but rather one in P600 amplitude (relative to a non-anomalous control), has challenged the core tenet in the electrophysiology of language that the N400 indexes processes of semantic integration and the P600 indexes syntactic processing. To explain such 'Semantic P600'-effects, while maintaining these views on the N400 and the P600, various *multi-stream* models of language processing have been devised 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 Depen-*

[1] Henceforth, we will use the terms 'Semantic Illusion'-effect and 'Semantic P600'-effect interchangeably.

dency Model: Bornkessel-Schlesewsky and Schlesewsky, 2008, and the *Processing Competition* account: Hagoort et al., 2009). All of these models include a processing stream that is purely semantic, unconstrained by any structural information (such as word order, agreement, and case marking, etc.). In processing a sentence such as 'De speer heeft de atleten geworpen' (lit: 'The javelin has the athletes thrown') this stream of independent semantic analysis does not encounter semantic processing problems on the word *thrown*, and hence does not produce an N400-effect (relative to a non-anomalous control 'De speer werd door de atleten geworpen'; lit: 'The javelin was by the athletes thrown'), because it can easily construct an interpretation on the basis of the words *javelin*, *athletes*, and *thrown*, which fit together rather well (i.e., the interpretation that the athletes have thrown the javelin). The output of this processing stream, however, conflicts with that from an algorithmic, structure-driven stream that does take surface structural information into account (i.e., producing the interpretation that the javelin has thrown the athletes). The effort put into resolving this problem, then, is reflected in a P600-effect, purportedly showing structural revision.

On the basis of a comprehensive review of multi-stream models and the empirical data at hand, Brouwer et al. (2012) conclude that none of the models is capable of explaining the full range of relevant findings in the literature. For one, most of the proposed models have difficulty explaining biphasic N400/P600 effects, such as those found by Hoeks et al. (2004) in response to 'De speer heeft de atleten opgesomd' (lit: 'The javelin has the athletes summarized', meaning that the javelin summarized the athletes) relative to a non-anomalous control 'De speer werd door de atleten geworpen' (lit: 'The javelin was by the athletes thrown'). Here, the words *javelin*, *athletes*, and *summarized* do not fit together well, and hence the independent semantic analysis stream should have difficulty constructing an interpretation, which should lead to an N400-effect. Critically, the algorithmic stream should agree with the independent semantic analysis stream that the sentence is infelicitous, meaning that the streams are not in conflict, and that there should be no P600-effect reflecting a conflict resolution process. This is inconsistent with the observed biphasic N400/P600-effect. What is more, none of the models can account for isolated P600-effects in a larger discourse. Nieuwland and van Berkum (2005), for instance, presented participants with short stories, like a story about a tourist checking into an airplane with a huge suitcase, and a woman behind the check-in counter deciding to charge the tourist extra because the suitcase is too heavy. This story was then continued with 'Next, the woman told the suitcase [...]'. At the word *suitcase*, all multi-stream models predict an N400-effect but no P600-effect (relative to the non-anomalous 'tourist'), because the independent semantic analysis stream and the algorithmic stream should agree upon the infelicity of the sentence so far. However, the reverse was actually found: the word *suitcase* produced a P600-effect, and no N400-effect (relative to *tourist*). Hence, these multi-stream models fall short of explaining the full breadth of relevant data.

In contrast to seeking a solution for the 'Semantic Illusion' phenomenon in aspects of cognitive architecture (e.g., an increase in number of processing streams), Brouwer et al. (2012) argued for a functional reinterpretation of the ERP components involved. First of all, in line with previous suggestions (Kutas and Federmeier, 2000; Lau et al., 2008; van Berkum, 2009), they propose that the N400 component reflects *retrieval* of lexical-semantic information, rather than semantic integration or any other kind of compositional semantic processing. Retrieving the information associated with a word is easier if that information is already (partially) activated by its prior context. This explains why the word *butter* engenders a much smaller N400 in the context of 'He spread his warm bread with […]' than the word *socks*. The lexical knowledge associated with *butter* is already activated by its prior context, as it fits well with *spread* and *warm bread*; in contrast, *socks* does not fit at all. It is important to note that this pre-activation stems from the preceding lexical items (*spread* and *bread*), as well as from the message representation that has been constructed so far (e.g., a breakfast scene). The retrieval view on the

N400 also explains the absence of an N400-effect in the ‘Semantic Illusion’ data: in both the target (‘The javelin has the athletes [. . .]’) and the control (‘The javelin was by the athletes [. . .]’) condition, the preceding context pre-activates the lexical features of an incoming word (e.g, *thrown*), yielding no difference in N400 amplitude, and hence no N400-effect.

If we accept the retrieval hypothesis on the N400, the question rises where in the ERP signal the integration of the retrieved word meaning into the unfolding utterance representation shows up. Brouwer et al. (2012) argue that these integrative processes are reflected in P600 amplitude. They hypothesize that the P600 is a family of late positive components, all of which reflect aspects of the word-by-word construction, reorganization, or updating of an utterance interpretation. This integrative processing may intensify (leading to an increase in P600 amplitude) in all kinds of processing circumstances, for instance, when new discourse entities need to be accommodated, relations between entities need to be established, thematic roles need to be assigned, information needs to be added to entities, already established relations need be revised, or when conflicts between information sources (e.g., with respect to world knowledge) need to be resolved. In other words, the compositional processes of integration and interpretation that were traditionally assumed to underlie N400 amplitude on the integration view, are hypothesized to be reflected in the amplitude of the P600 instead. This naturally explains the presence of a P600-effect in the ‘Semantic P600’ sentences: integrating the meaning of the critical word (*thrown*) with its prior context leads to an anomalous interpretation in the target condition (‘The javelin has the athletes [. . .]’), but not in the control condition (‘The javelin was by the athletes [. . .]’). Interestingly, this integration view on the P600 predicts that semantic anomalies, like “He spread his warm bread with socks/butter”, should not only produce an increase in N400 amplitude, but also an increase in P600 amplitude (see section 5.4 for further discussion). A close look at the data reveals that this is indeed the case (see Kutas and Hillyard, 1980, Fig. 1c, which clearly reveals a biphasic N400/P600 pattern). Not only semantic anomalies, but also syntactic complexities and anomalies can elicit a P600-effect. On the Integration view on the P600, these effects reflect difficulties in establishing a coherent utterance representation, rather than processes operating on a purely syntactic representation (Brouwer et al., 2012). We will return to this issue in the discussion.

Given the Retrieval view on the N400 component and the Integration view on the P600 component, Brouwer et al. (2012) suggested that language is processed in biphasic N400/P600—Retrieval–Integration—cycles: Every incoming word modulates the N400 component, reflecting the processes involved in activating its associated conceptual knowledge. Every word also modulates the P600 component, reflecting the processes involved in integrating this activated knowledge into an updated utterance representation. Although they have argued that the resultant single stream Retrieval–Integration (RI) account has the broadest empirical coverage of extant models (Brouwer et al., 2012; Hoeks and Brouwer, 2014), this account is still a conceptual one. In what follows, we will offer a formally precise neurocomputational instantiation of the RI account, and present a simulation of the results of an ERP experiment by Hoeks et al. (2004), thereby providing the RI account with a proof of concept.

3 The neurocomputational model

Our aim is to derive a neurocomputational model of language processing that indexes the N400 and the P600 components of the ERP signal in semantic processing. In deriving such a model, we want to minimally adhere to the following design principles. First, we should model the N400 and the P600 in a single comprehension architecture, rather than in two separate models. Second, we should model the right level of granularity: We aim to index scalp-recorded summations of post-synaptic potentials in large neural populations, and we should therefore model the processes underlying the N400 and the

Table 1: Materials and effects of the Hoeks et al. study. Materials used in the ERP experiment by Hoeks et al. (2004), as well as the effects that were observed for each condition (relative to control).

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

P600 at that level. Third, estimates of N400 and P600 amplitude should emerge from the processing behavior of the model, that is, the model should not be explicitly trained to produce these estimates. Fourth, the model should account for the relevant patterns of ERP-effects induced by semantic processing. We take the model to be successful if for a given contrast it produces the correct N400-effect and/or P600-effect (or the absence thereof). Finally, to assure the generalizability of the model, we want to obtain these effects in at least two separate, independent simulations.

In what follows, we will show how we derived a neurocomputational model of language processing that adheres to these design principles. To satisfy the first three principles, we employ a single artificial neural network architecture, which offers the right level of granularity and allows for modeling the N400 and the P600 as emergent epiphenomena of processing. To satisfy principle four, we want to capture 'Semantic P600'-effects, as well as traditional semantically-induced biphasic N400/P600-effects (cf. Kutas and Hillyard, 1980). To this end, we will model an experiment by Hoeks et al. (2004). 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 revealed a P600-effect on the final verb *thrown*, but no N400-effect (see Table 1 and Fig. 1). Two other semantically anomalous conditions were also compared to the

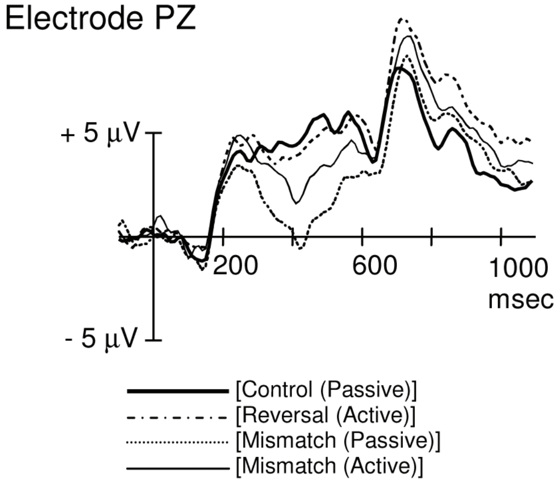

Fig. 1: Results of the Hoeks et al. study. Results (Pz electrode) of the ERP experiment by Hoeks et al. (2004). Positive is plotted upwards. Note that this single electrode only serves as an illustration; our present simulation results are compared to the (statistically evaluated) effects found on the whole array of electrodes used in the original study.

same control, and both showed 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') (see Table 1 and Fig. 1). Finally, to satisfy principle five, we will conduct two independent simulations of this experiment. Below, we will first introduce the overall architecture of the model, which we will subsequently break down in detail. Next, we will introduce the results of our simulations of the Hoeks et al. experiment.

3.1 Model architecture

The RI account postulates that incremental, word-by-word language processing proceeds in Retrieval–Integration cycles. Mechanistically, each cycle can be conceptualized as a function $process : (\textit{word form}, \textit{utterance context}) \to \textit{utterance representation}$, mapping a perceived word w_t (word form), and the context established after processing words $w_1 \ldots w_{t-1}$ that precede w_t (utterance context), into an updated utterance interpretation spanning words $w_1 \dots w_t$ (utterance representation). Critically, this mapping is not direct. The RI account breaks it down into a Retrieval and an Integration sub-process by introducing an intermediate representation encoding the meaning of word w_t. This intermediate representation is the output of the Retrieval process, which can be conceptualized as a function $retrieve$: $(\textit{word form}, \textit{utterance context}) \rightarrow \textit{word meaning}$, mapping a word w_t (word form) into a representation of its meaning (word meaning), while taking into account the context in which it occurs (utterance context). The output of this Retrieval process, a representation of the meaning of word w_t, serves as input to the Integration process. During Integration, it is combined with the context established after processing words $w_1 \dots w_{t-1}$ to produce an updated utterance representation spanning the entire utterance $w_1 \dots w_t$. Hence, the Integration process can be conceptualized as a function $integrate$: $(\textit{word meaning}, \textit{utterance context}) \rightarrow \textit{utterance representation}$, mapping the meaning of a word w_t (word meaning) and its prior context (utterance context) into an updated utterance interpretation (utterance representation).

Our neurocomputational model is effectively an 'extended' Simple Recurrent Network (SRN; Elman, 1990) that instantiates the $process$ function, broken down into its $retrieve$ and $integrate$ sub-processes. Fig. 2 provides a schematic overview of our model. The model consists of five layers of artificial neurons, implementing the input to the model (INPUT), a Retrieval module (RETRIEVAL and RETRIEVAL_OUTPUT), and an Integration module (INTEGRATION and INTEGRATION_OUTPUT). All artificial neurons in the model are logistic dot-product units, meaning that the activation level y_j of a unit j is defined as:

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

where x_j is the net input to unit j:

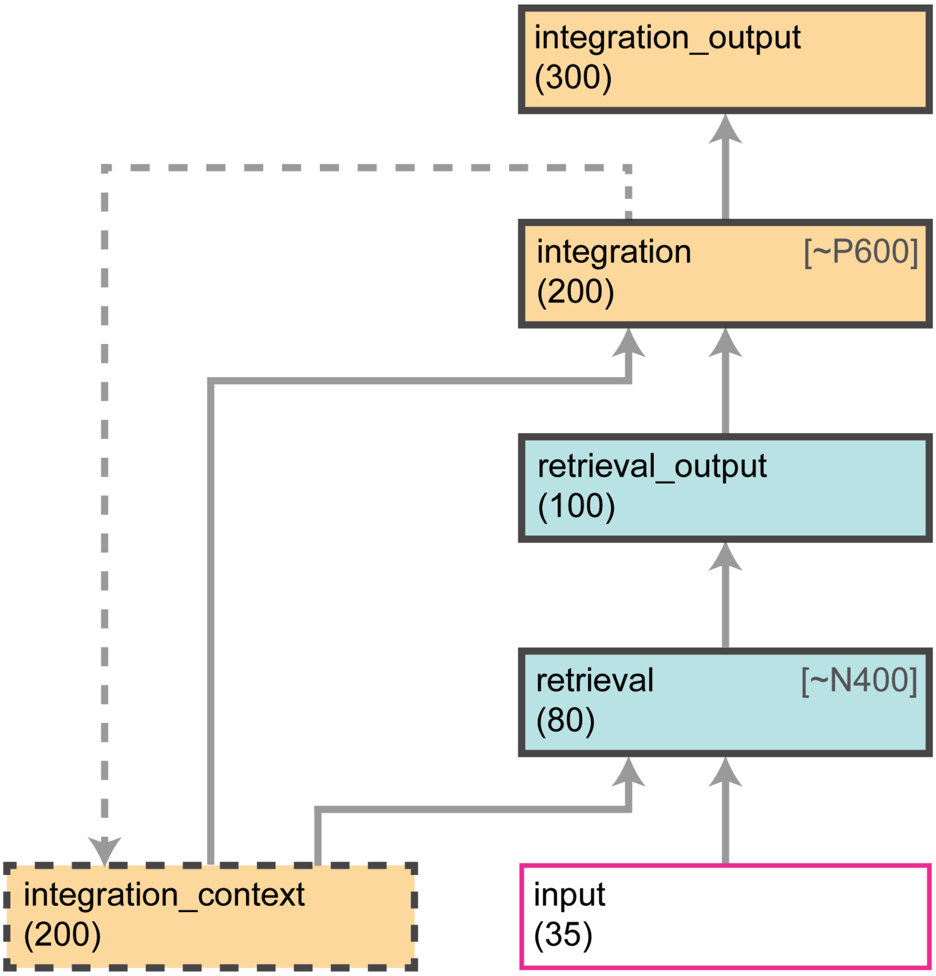

Fig. 2: The neurocomputational model. Schematic illustration of the neurocomputational model instantiating the Retrieval–Integration account of the N400 and the P600. Each rectangle represents a vector of artificial (logistic dot-product) neurons, and each solid arrow represents a matrix of connection weights that connect each neuron in a projecting layer to each neuron in a receiving layer. The network receives its input at the INPUT layer, and produces its output at the INTEGRATION_OUTPUT layer. The rectangle with a dashed border is a context layer (cf. Elman, 1990), and the dashed arrow represents a copy projection; prior to feedforward propagation of activation from the INPUT to the INTEGRATION_OUTPUT layer, the INTEGRATION_CONTEXT layer receives a copy of the INTEGRATION layer. At time-step $t = 0$, the activation value of each unit in the INTEGRATION_CONTEXT layer is set to 0.5. All layers except the INPUT and INTEGRATION_CONTEXT layer also receive input from a bias unit (not shown), the activation value of which is always 1.

$$x_j = \sum_i y_i w_{ij} \tag{2}$$

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 from i to j. Time in the model is discrete, and at each processing timestep t, activation flows from the INPUT layer, through the RETRIEVAL layer to the RETRIEVAL_OUTPUT layer, and from RETRIEVAL_OUTPUT layer through the INTEGRATION layer to the INTEGRATION_OUTPUT layer. To allow for context-sensitive retrieval and integration (see below), the RETRIEVAL and the INTEGRATION layer both also receive input from the activation pattern in the INTEGRATION layer as established at the previous timestep $t-1$, effectuated through an additional context layer (INTEGRATION_CONTEXT; cf. Elman, 1990). Prior to feedforward propagation of activation from the INPUT to the INTEGRATION_OUTPUT layer, this INTEGRATION_CONTEXT layer receives a copy of the INTEGRATION layer (at time-step $t = 0$, the activation value of each unit in the INTEGRATION_CONTEXT layer is set to 0.5). Finally, all layers except the INPUT and INTEGRATION_CONTEXT layer also receive input from a bias unit, the activation value of which is always 1.

Overall, the model is trained to map sequences of words forms, clamped onto the INPUT layer, into an utterance representation encoding sentence meaning in the INTEGRATION_OUTPUT layer. It does so on an incremental, word-by-word basis, thereby instantiating the $process$ function (the processing of a word takes a single time tick; cf. Elman, 1990). In the model, word forms are localist word representations, and the utterance representations are thematic-role assignment representations (i.e, *who*-does-*what*-to-*whom/what*; cf. Mayberry et al., 2009; Crocker et al., 2010). We will further elaborate upon these representations below. Importantly, the mapping from word forms into an utterance representation is not direct, as it is broken down into the $retrieve$ and $integrate$ sub-processes, which can be directly linked to the N400 and the P600 component, respectively. Provided an incoming word w_t (INPUT), and the unfolding context

(INTEGRATION_CONTEXT), the RETRIEVAL layer serves to activate a word meaning representation of w_t in the RETRIEVAL_OUTPUT layer. Hence, the function of the RETRIEVAL layer is to $retrieve$ word meaning. In the model, word meaning representations take the form of distributed semantic feature vectors, which we will elaborate below. The INTEGRATION layer, in turn, combines the activated word meaning representation (RETRIEVAL_OUTPUT) with the unfolding context (INTEGRATION_CONTEXT), into an updated utterance representation (INTEGRATION_OUTPUT). The INTEGRATION layer thus serves to $integrate$ word meaning into the unfolding interpretation.

In what follows, we will provide a detailed derivation of the overall model, broken down into a Retrieval and an Integration module. We will describe how each of these modules was trained, and we will give a formal description of the representations involved. Next, we will provide a word-by-word walk-through of the processing dynamics of the model, as well as a derivation of the linking hypotheses to the N400 and the P600 component.

3.2 (De)constructing the Integration module

In the model, the processing of a single word entails two context-sensitive mappings: one from a word form representation into a word meaning representation ($retrieve$), and one from a word meaning representation into an utterance representation ($integrate$). In order to get the model to produce the intermediate word meaning representations, we employ a two-stage training procedure in which we first construct the Integration module, and subsequently add the Retrieval module (the reason for this will be explained in section 3.3.2).

The Integration module is a sub-network of the overall model: An SRN consisting of the RETRIEVAL_OUTPUT (input to the SRN), INTEGRATION (hidden), INTEGRATION_OUTPUT (output), and INTEGRATION_CONTEXT (context) layers (hence, it is the overall model minus the INPUT and RETRIEVAL layers). Recall that we want the Integration module to implement the function $integrate : (\textit{word meaning}, \textit{utterance context}) \rightarrow$ *utterance representation*, mapping the meaning of a word w_t (word meaning) and its prior context (utterance context) into an updated utterance interpretation (utterance representation). To this end, we train the Integration module to map sequences of word meaning representations, clamped onto the RETRIEVAL_OUTPUT layer, into an utterance representation in the INTEGRATION_OUTPUT layer.

3.2.1 Word meaning representations

Following the intuition behind many influential theories of word meaning, our model employs feature-based semantic representations as word meaning representations (see McRae et al., 2005, for an overview on semantic features). More specifically, it employs 100-dimensional binary representations, which were derived from a large corpus of Dutch newspaper texts (the TwNC corpus; Ordelman et al., 2007) using the Correlated Occurrence Analogue to Lexical Semantics (COALS; Rohde et al., 2009). We derived these representations both for words belonging to the open word class as well as for words belonging to the closed word class. Appendix B provides a detailed description of this derivation procedure.

3.2.2 Utterance representations

The utterance representations produced by our model are thematic-role assignment representations (i.e, *who*-does-*what*-to-*whom/what*; cf. Mayberry et al., 2009; Crocker et al., 2010). These thematic-role assignment representations are 300-dimensional vectors, which are divided into three 100-dimensional slots. These three slots respectively identify the word meaning representations of the elements that will be *agent*, *action*, and *patient* (cf. Mayberry et al., 2009).

3.2.3 Training data

Our aim is to model the study by Hoeks et al. (2004), which includes active and passive constructions (see Table 1). To this end, we want the model to 'understand' sentences with the following template structure:

Active sentences:

de	[Agent]	heeft	het/de	[patient]	[action]
the	[agent]	has	the$_{(+/-\text{NEUTER})}$	[patient]	[action]

passive sentences:

het/de	[patient]	werd	door	de	[agent]	[action]
the$_{(+/-\text{NEUTER})}$	[patient]	was	by	the	[agent]	[action]

importantly, we want the model to 'know' (like human language users) that 1) any noun can theoretically be an *agent* or a *patient* (productivity), but 2) there are certain combinations of *agents*, *actions*, and *patients* that are more plausible (stereotypicality; $\sim$minimal world knowledge, cf. Mayberry et al., 2009). for each simulation, we generated a separate set of training sentences from the above templates, by filling in the *agent*, *patient*, and *action* slots using the nouns (agent and patient) and verbs (action) listed in table 2 (note that each of the two simulations uses a completely different set of nouns and verbs, corresponding to different word meaning representations, and therefore different utterance representations). Agent nouns were always preceded by the determiner 'de' (the$_{-\text{NEUTER}}$), and patient nouns by either 'de' (the$_{-\text{NEUTER}}$) or 'het' (the$_{+\text{NEUTER}}$), depending on the gender of the noun (note that 'het' unambiguously signals a neuter noun in Dutch; see Table 2). The items in each set can be divided into two halves. The first half of each set is intended to teach the model principle (1) and consists of sentences constructed by permuting each of the twenty nouns (agents plus patients) with each verb, yielding $20 \times 20 \times 10 = 4000$ active sentences, and $20 \times 20 \times 10 = 4000$ passive sentences (i.e., 8000 items in total). The second half of training set is intended to induce principle (2) and consists only of sentences with stereotypical agent-action-patient combinations (rows of Table 2). This means that each stereotypical triplet occurs $8000/10 = 800$ times in this half of the training data. Again, half of these items are actives, and half of them are passives. As a result, each full set contains 16000 training items (50% actives and 50% passives), in which each verb appears 1600 times, 802 times ($\approx 1 : 1$) of which in a stereotypical agent-action-patient construction, and 2 times ($\approx 0.001\%$) of which in *each* of the non-stereotypical constructions. Hence, each stereotypical agent-action-patient construction occurs 401 times more frequently than any non-stereotypical agent-action-patient triplet. Overall, this yields a stereotypicality/non-stereotypicality ratio of $.50125/.49875 \approx 50\%$, which we take to reflect no a priori bias towards either productivity (principle 1) or stereotypicality (principle 2) (cf. Mayberry et al., 2009).

As the Integration module must learn to process 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 word meaning representations, and the target is always the desired utterance representation. Note that for anomalous agent-action-patient combinations, these targets also reflect the corresponding anomalous utterance representations.

3.2.4 Training procedure

We trained the Integration module using bounded gradient descent (Rohde, 2002), a modification of the standard backpropagation algorithm (Rumelhart et al., 1986). Each model (i.e., one for each simulation) was trained for 7000 epochs, minimizing Mean Squared Error (MSE). In each epoch, gradients were accumulated over 100 items before updating the weights (within each item, error was backpropagated after each word). 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 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 was set to a constant of 0.9. Finally, we used a zero error radius of 0.1, such that no error was backpropagated when the difference between the produced activity level y_j of a unit j and the desired activity level d_j of this unit was smaller than 0.1, that is, when $|y_j - d_j| < 0.1$. Appendix C provides a detailed, mathematical description of the training procedure.

After training, we evaluated the comprehension performance of the model using an 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 (see Appendix C). 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. Comprehension performance was perfect (100% correct) for each of the two models ($MSE_{model_1} = 0.212$; $MSE_{model_2} = 0.206$).

3.3 (De)constructing the Retrieval module

With the Integration module in place, we can now add in the Retrieval module to arrive at the overall model as outlined above (and depicted in Fig. 2). Like the Integration module, the Retrieval module can be seen as a sub-network of the overall model: An SRN consisting of the INPUT (input to the SRN), RETRIEVAL (hidden), RETRIEVAL_OUTPUT (output), and INTEGRATION_CONTEXT (context) layers. Recall that we want the Retrieval module to implement the function $retrieve : (\textit{word form}, \textit{utterance context}) \rightarrow \textit{word meaning}$, mapping a word w_t (word form) into a representation of its meaning (word meaning), while taking into account the context in which it occurs (utterance context). To this end, we train the Retrieval module to map word form representations, clamped onto the INPUT layer, into word meaning representations in the RETRIEVAL_OUTPUT layer, while taking into account the unfolding context in the INTEGRATION_CONTEXT layer.

3.3.1 Word form representations

The word form representations that serve as input to the Retrieval module, and hence to the overall model, are localist word representations encoding word identity. That is, the model employs 35-dimensional localist word representations, in which each unit corresponds to a single word (20 nouns + 10 verbs + 2 auxiliary verbs + 2 determiners + 1 preposition = 35 words).

3.3.2 Training data and procedure

If one ignores context (INTEGRATION_CONTEXT), the mapping from word form representations (INPUT) into word meaning representations (RETRIEVAL_OUTPUT), the $retrieve$ function, is a straightforward recoding problem; the RETRIEVAL layer must simply map each of the 35 unique word form representations into its corresponding, unique word meaning representation. However, on the RI account, the retrieval of word meaning is assumed to be strongly context-driven, and therefore we want the RETRIEVAL layer to take into account the utterance context in the INTEGRATION_CONTEXT layer. For reasons laid out below, getting the model to produce this behavior is not straightforward, and in order to obtain the intended behavior, we derive a rather non-standard training procedure, which involves training the Retrieval module as part of the overall model. Note, however, that we do not posit our model as a model of language acquisition, and hence do not attribute any psychological or biological reality to our training procedure; our only interest is the resultant comprehension model.

An intuitively attractive, but incorrect approach would be to train the Retrieval module as a standalone SRN, before combining it with the Integration module to arrive at the overall model. That is, one could present the Retrieval module with sequences of input-target patterns, of which the inputs are word form representations and the targets are word meaning representations. This approach requires the utterance contexts as constructed by the Integration module (integration_context), which could in principle be 'recorded', and presented to the Retrieval module as inputs along with the word form representations. However, this approach boils down to the same straightforward recoding outlined above (i.e., mapping from 35 unique word form representations into their corresponding, unique word meaning representations), but now with an additional source of information: the utterance contexts. Because these contexts vary substantially across sentences, they will be nothing but noise to the model, and hence the model is better off ignoring them. Crucially, this is true for all approaches in which the Retrieval module is explicitly trained to produce the correct word meaning representations (see Appendix D for empirical support). Hence, we need an approach to training that pressures the model to take utterance context into account. To achieve this, we trained the Retrieval module as part of the overall network. After training the Integration module, we added in the input, and retrieval layers, to arrive at the overall model as depicted in Fig. 2. In this model, we froze all the weights in the Integration module, that is, all weights on the projections: retrieval_output → integration, integration → integration_output, and integration_context → integration (as well as those on the relevant biases). We then trained the overall model using the same procedure and the same patterns as we used for training the Integration module, with the exception that the inputs to the model were now word form representations, clamped onto the input layer. Thus, the overall model has to learn to map sequences of word form representations into an utterance representation. This approach has two important consequences for our desired behavior.

First, due to the fact that the Integration module is no longer malleable (its weights are frozen), it becomes fully deterministic. This means that in order to map a sequence of word form representations (input) into an utterance representation (integration_output), the model has to find a way to activate the right inputs to the Integration module (RETRIEVAL_OUTPUT). As the Integration module was trained to high accuracy, this means that for a given word form representation, the output of the Retrieval module (RETRIEVAL_OUTPUT) is forced to be close to the corresponding word meaning representation (or a subset of its relevant semantic features, depending on the solution found during the training of the Integration module).

Second, the error signal is now driven by the desired utterance representations, rather than by the word meaning representations. This means that in order for the model to incrementally construct such an utterance representation, it must take into account the utterance context. That is, without context, it would simply try to map each individual word form (in isolation) into an utterance representation. Hence, this approach pressures the Retrieval module of the model to take the utterance context (INTEGRATION_CONTEXT) into account when trying to map a word form (INPUT) into the utterance representation (INTEGRATION_OUTPUT). As described above, this in itself requires the model to activate the right inputs (RETRIEVAL_OUTPUT)—the right word meaning representations (or relevant semantic features thereof)—for the Integration module.

After training, we again computed an output-target similarity matrix, but this time on the overall model. Comprehension performance was perfect (100% correct) for each of the two models ($MSE_{model_1} = 0.273$; $MSE_{model_2} = 0.248$). Moreover, we also computed the cosine similarity between the produced word-meaning representations at the RETRIEVAL_OUTPUT and their desired targets, for each word in each possible sentence. As expected, the Retrieval module outputs word meaning representations that are highly similar to the word meaning representations used in training the Integration module ($cos_{model_1} = .951$ (sd=.023); $cos_{model_2} = .961$ (.021)).

3.4 Processing in the model

Now that we have arrived at the full model, we can walk through its processing dynamics on a word-by-word basis. As our aim is to model the ERP experiment by Hoeks et al. (2004), we will show how the model processes each of the four conditions contrasted in this study (see Table 1). To this end, we constructed an example sentence for each condition, using materials from simulation 1: a control sentence 'De maaltijd werd door de kok bereid' (lit: 'The meal was by the cook prepared'; [Control (Passive)]), a role-reversed sentence 'De maaltijd heeft de kok bereid' (lit: 'The meal has the cook prepared'; [Reversal (Active)]), a passive mismatch sentence 'De maaltijd werd door de kok gezongen' (lit: 'The meal was by the cook sung'; [Mismatch (Passive)]), and an active mismatch sentence 'De maaltijd heeft de kok gezongen' (lit: 'The meal has the cook sung'; [Mismatch (Active)])—see section 4 for more details on how we derived these sentences. To gain insight into what the model anticipates at different points in processing, Fig. 3 shows how close (in terms of cosine similarity) the word meaning representation in each thematic-role slot (in the INTEGRATION_OUTPUT layer) is to either the representation of each of the nouns ('kok'/'cook' and 'maaltijd'/''meal') for the *agent* and *patient* slots, or to that of each of the verbs ('bereid'/'prepared' and 'gezongen'/'sung') for the *action* slot.

A first thing to note about the processing dynamics of the model is that once it encounters the first noun 'maaltijd'/'meal', it moves from a state of relative indecision about the sentence interpretation (at 'de'/'the')[2], to a state in which it strongly anticipates the interpretation that the 'meal was prepared by a cook'; hence not only does it anticipate 'meal' to obtain a *patient* role, it also anticipates (albeit to a somewhat lesser extent) 'cook' to be the *agent* and 'prepared' to be the *action*. If the sentence unfolds as passive construction ([Control (Passive)] and [Mismatch (Passive)] conditions in Fig. 3), signalled by the auxiliary verb 'werd'/'was', these predictions are gradually confirmed by the consecutive words. The sentence-final verb, then, either completely

[2] The difference in anticipation about the *agent* role after processing the determiner 'de'/'the$_{-\text{NEUTER}}$' is due to the fact that this determiner occurs less frequently with inanimate nouns than the neutral determiner 'het'/'the$_{+\text{NEUTER}}$', whereas all animate nouns occur with 'de'/'the$_{-\text{NEUTER}}$' (see Table 2). Since animate nouns typically occur in the *agent* role in the training data, encountering a sentence-initial 'de'/'the$_{-\text{NEUTER}}$' results in a slightly biased anticipation towards 'kok'/'cook' relative to 'maaltijd'/'meal' as *agent*.

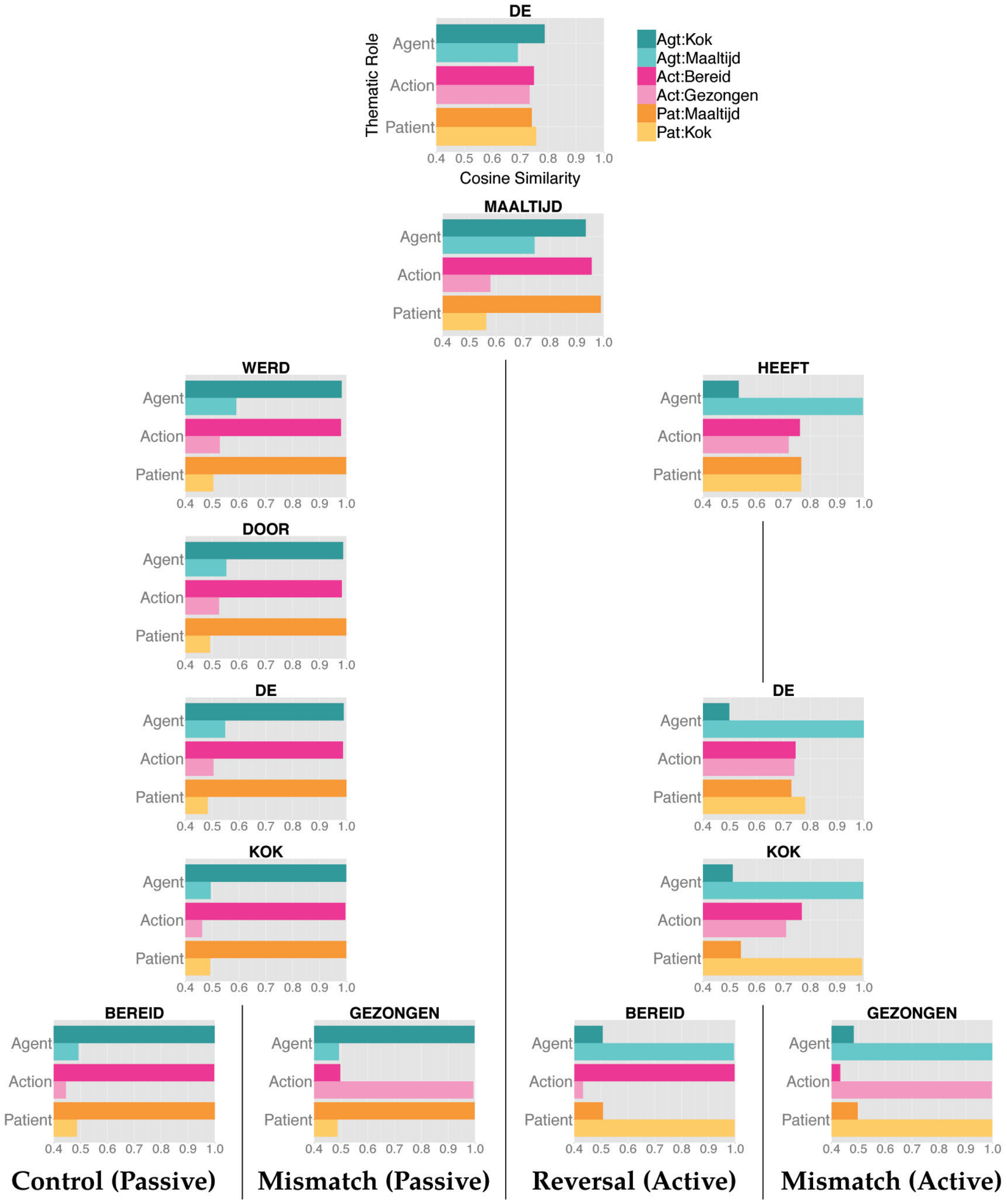

Fig. 3: Word-by-word walk-through of processing in the model. Illustration of the word-by-word processing of an example sentence (from simulation 1) for each condition of the Hoeks et al. (2004) experiment (see text). The bar plots show the cosine similarity of the word meaning representation in each thematic-role slot (in the INTEGRATION_OUTPUT layer) relative to either the representation of each of the nouns ('kok'/'cook' and 'maaltijd'/''meal') for the *agent* and *patient* slots, or to that of each of the verbs ('bereid'/'prepared' and 'gezongen'/'sung') for the *action* slot.

confirms the anticipated interpretation ('bereid'/'prepared') or disconfirms it by signalling that the anticipated *action* should be revised ('gezongen'/'sung'). On the other hand, if the noun 'maaltijd'/'meal' turns out to be part of an active construction ([Reversal (Active)] and [Mismatch (Active)] conditions in Fig. 3), as signalled by the auxiliary 'heeft'/'has', the model immediately revises its anticipated interpretation to one in which the 'maaltijd'/'meal' is assigned the role of *agent*, and in which the *patient* and *action* have yet to be determined. The model updates the former upon encountering the second noun 'kok'/'cook', and the latter upon encountering the sentence-final verb ('bereid'/'prepared' or 'gezongen'/'sung').

Crucially, the model thus differentially anticipates the interpretations of active and passive sentences, as well as the sentence-final verbs across these constructions (i.e., see the interpretation constructed at the pre-final word 'kok'/'cook'). This means that the internal representation at the INTEGRATION layer of the model contains different information across these constructions. Consequently, the contextual inputs to the RETRIEVAL layer (from the INTEGRATION_CONTEXT → RETRIEVAL projection) and the INTEGRATION layer of the model (from the INTEGRATION_CONTEXT → INTEGRATION projection) also differ across these constructions. If we compare the activation pattern in the INTEGRATION layer after processing the noun 'kok'/'cook' in either an active or a passive construction, using cosine similarity, these patterns are clearly different (cos: .569). Moreover, if we look at the activation patterns in the RETRIEVAL layer and the INTEGRATION layer after processing the sentence-final verb 'bereid'/'prepared', we see an effect of these different contexts. That is, although the sentence final words (and therefore their word form and word meaning representations) are the same, the activation patterns are different at the RETRIEVAL layer (.847) and the INTEGRATION layer (.567) across actives and passives. Hence, the activity patterns at both the RETRIEVAL and the INTEGRATION layer are modulated by context. Interestingly, the effect of context is not reflected in the output of Retrieval module at the RETRIEVAL_OUTPUT layer (.952); the same word forms obtain highly similar word meaning representations in different contexts (as is evidenced by the mean similarity scores reported at the end of section 3.3.2). If, on the other hand, different sentence final-words (e.g., 'bereid'/'prepared' versus 'gezongen'/'sung') occur in the same context, this also modulates the activation patterns at the RETRIEVAL (active: .359; passive: .268) and INTEGRATION layer (active: .735; passive: .771). In summary, then, the RETRIEVAL and INTEGRATION layers appear to successfully implement the $retrieve$ and $integrate$ functions, respectively.

3.5 Linking hypotheses

Provided the observed processing behavior of the model, we can now formulate a linking hypothesis between N400 amplitude and activity in the RETRIEVAL layer, and a linking hypothesis between P600 amplitude and activity in the INTEGRATION layer.

3.5.1 Linking hypothesis to the N400

On the RI account, N400 amplitude is an index of the amount of processing involved in activating the conceptual knowledge associated with an incoming word in memory. More specifically, at any given point in processing, we assume the semantic memory system to be in a particular state, reflecting the preceding word and prior context. Upon encountering a next word in the current context, activation of the conceptual knowledge associated with this word involves an alteration of this state. The process of altering the state of semantic memory from one word to the next is what we assume to be reflected in the N400 component; if the previous and new state are relatively similar (because the new state was anticipated by the previous state; i.e., context pre-activated the conceptual knowledge associated with the incoming word), state transition requires little work, and N400 amplitude will be reduced. If, on the other hand, the previous and new state are highly dissimilar (context did non pre-activate the conceptual knowledge associated with the incoming word), state transition requires more effort, and N400 amplitude is increased. Hence, we take N400 amplitude to be a measure of the *processing* induced by a mismatch between the predicted conceptual knowledge and the conceptual knowledge associated with the observed word; that is, we do not take N400 amplitude to be a direct measure of the mismatch itself.

In the model, retrieval processes take place in the RETRIEVAL layer, which implements the function $retrieve$: (*word form*, *utterance context*) $\rightarrow$ *word meaning*. Given the identity (word form) of a given word w_t, and the utterance context as established after processing words $w_1 \ldots w_{t-1}$ that precede w_t (INTEGRATION_CONTEXT), the Retrieval module will draw upon is input history to activate the meaning representation (or the relevant semantic features thereof) corresponding to word w_t. Crucially, the training of the Retrieval module was driven by the utterance representation, which forced the model into a context-dependent solution for this word form to word meaning mapping. As result, the internal representations constructed at the RETRIEVAL (hidden) layer of the module are high-dimensional abstractions over the word form representations and utterance contexts (its inputs) and word meaning representations (its outputs), rather than intermediate word meaning representations. Hence, any changes in the activation pattern of the RETRIEVAL layer can be taken to reflect changes in the semantic memory state of the model, and therefore as processing required for activating and/or deactivating semantic features. As such, we estimate N400 amplitude for a given word w_t as the degree of change that this word induces in the activity pattern of the RETRIEVAL layer, provided the activity pattern as established after processing the previous word w_{t-1}, using cosine dissimilarity:

$$\text{N400} = 1 - cos(\textsc{retrieval}_t, \textsc{retrieval}_{t-1}) \tag{3}$$

3.5.2 Linking hypothesis to the P600

On the RI account, P600 amplitude reflects the amount of processing involved in the word-by-word construction, reorganization, or updating of an utterance interpretation. In the model, this processing takes place in the INTEGRATION layer, which implements the *integrate* function. Given the meaning of a word w_t (RETRIEVAL_OUTPUT), and the utterance context as established after processing words $w_1 \ldots w_{t-1}$ that precede w_t (INTEGRATION_CONTEXT), the Integration module will draw upon its input history to predict the most likely utterance representation for the sentence so far (see section 3.4). Crucially, the anticipatory state of the Integration module (an SRN) is contained within its internal representation at the INTEGRATION (hidden) layer, which constitutes a high-dimensional abstraction over the inputs to the module (word meaning representations and utterance contexts), and its outputs (utterance representations). Hence, any changes in the activation pattern of the INTEGRATION layer can be taken to reflect processing involved in (re)composing the utterance representation. As such, we estimate P600 amplitude for a given word w_t as the degree of change that its meaning induces in the activity pattern of the INTEGRATION layer, provided the activity pattern as established after processing the previous word w_{t-1}, using cosine dissimilarity (cf. Crocker et al., 2010):

$$\text{P600} = 1 - cos(\text{INTEGRATION}_t, \text{INTEGRATION}_{t-1}) \tag{4}$$

4 Simulations

With the model in place, we can now turn to the simulation of the actual data from an ERP experiment. Recall that we want our model to capture 'Semantic P600'-effects, as well as traditional semantically-induced biphasic N400/P600-effects (cf. Kutas and Hillyard, 1980). To this end, we set out to model the study by Hoeks et al. (2004). Table 1 lists the materials of this study and their associated effects, and Fig. 1 shows their ERP modulations at the Pz electrode.

4.1 Testing procedure

To test the contrasts from the Hoeks et al. (2004) study in the model, we generated two sets of 40 test sentences, one set for each simulation. Each set contains 10 passive stereotypical agent-action-patient sentences [Control (Passive)], 10 active role-reversed sentences [Reversal (Active)], 10 passive semantic mismatch sentences [Mismatch (Passive)], and 10 active semantic mismatch sentences [Mismatch (Active)]. The role-reversed sentences were constructed by swapping the stereotypical agents and patients. The passive mismatch sentences were constructed in the same way as the control sentences, except that the stereotypical action verb was replaced by a mismatch verb (listed in Table 2). The active mismatch sentences, finally, were constructed like the role-reversed sentences, but also had the stereotypical action verb replaced by a mismatch verb.

We presented these materials to the model[3], and recorded its N400 and P600 estimates at the critical words. The model produces estimates in the range $[0, 1]$ (cosine dissimilarities), whereas the original ERP modulations are voltage fluctuations on an 'unbounded' microvolts scale $[-\mu V, +\mu V]$. In order to visually compare the model estimates to the ERP data, we therefore transformed the ERP amplitudes to a zero-to-one scale, using the transformation $(\mu V - min(\mu V))/(max(\mu V) - min(\mu V))$. Fig. 4 compares the N400 estimates as produced by the models (one for each simulation) to the N400 modulations in the ERP experiment (at the Pz electrode), and Fig. 5 shows the comparison between the P600 estimates produced by the models, and the P600 modulations in the ERP experiments (also at the Pz electrode). Note that this visual comparison of the model estimates to the ERP modulations at the Pz electrode only serves as an il-

[3] Note that the test items are a subset of the training items, and that comprehension performance is therefore perfect (100% correct) on both test sets.

lustration. Below, we will compare our simulation results to the (statistically evaluated) effects found on the whole array of electrodes used in the original study.

4.2 N400 results

Recall that we take the model to be successful if for a given contrast it produces the same N400-effect and P600-effect (or the absence thereof) as found in the Hoeks et al. (2004) study (see Table 1). For the N400, statistical evaluation using Repeated Measures ANOVA (with Condition as four-level within-items factor and Huynh-Feldt correction where necessary) showed a successful simulation of the original findings. The main effect of Condition was significant in each of the simulation experiments (Exp 1: F(3,27)=45.1; p<.001; Exp 2: F(3,27)=12.3; p<.001), and subsequent pairwise comparisons showed that 1) the N400-effect was absent in role-reversed 'Semantic Illusion' sentences (Exp 1: p=.47 [Bonf.]/p=.08 [Uncorr.]; Exp 2: p=.91 [Bonf.]/p=.15 [Uncorr.]), and 2) there was a significant N400-effect for the two other anomalous conditions (Exp 1: p-values<.005; Exp 2: p-values<.01).

4.3 P600 results

For the P600 component, we also successfully simulated the original findings. In both simulations, P600 amplitude was significantly higher for all anomalous sentences (including the role reversed 'Semantic Illusion' sentences) compared to controls (Main effect of Condition: Exp 1: F(3,27)=136.5; p<.001; Exp 2: F(3,27)=70.1; p<.001); pairwise comparisons showed that there was a significant P600-effect for all three anomalous conditions compared to control (Exp 1: all three p-values<.001; Exp 2: all three p-values<.001).

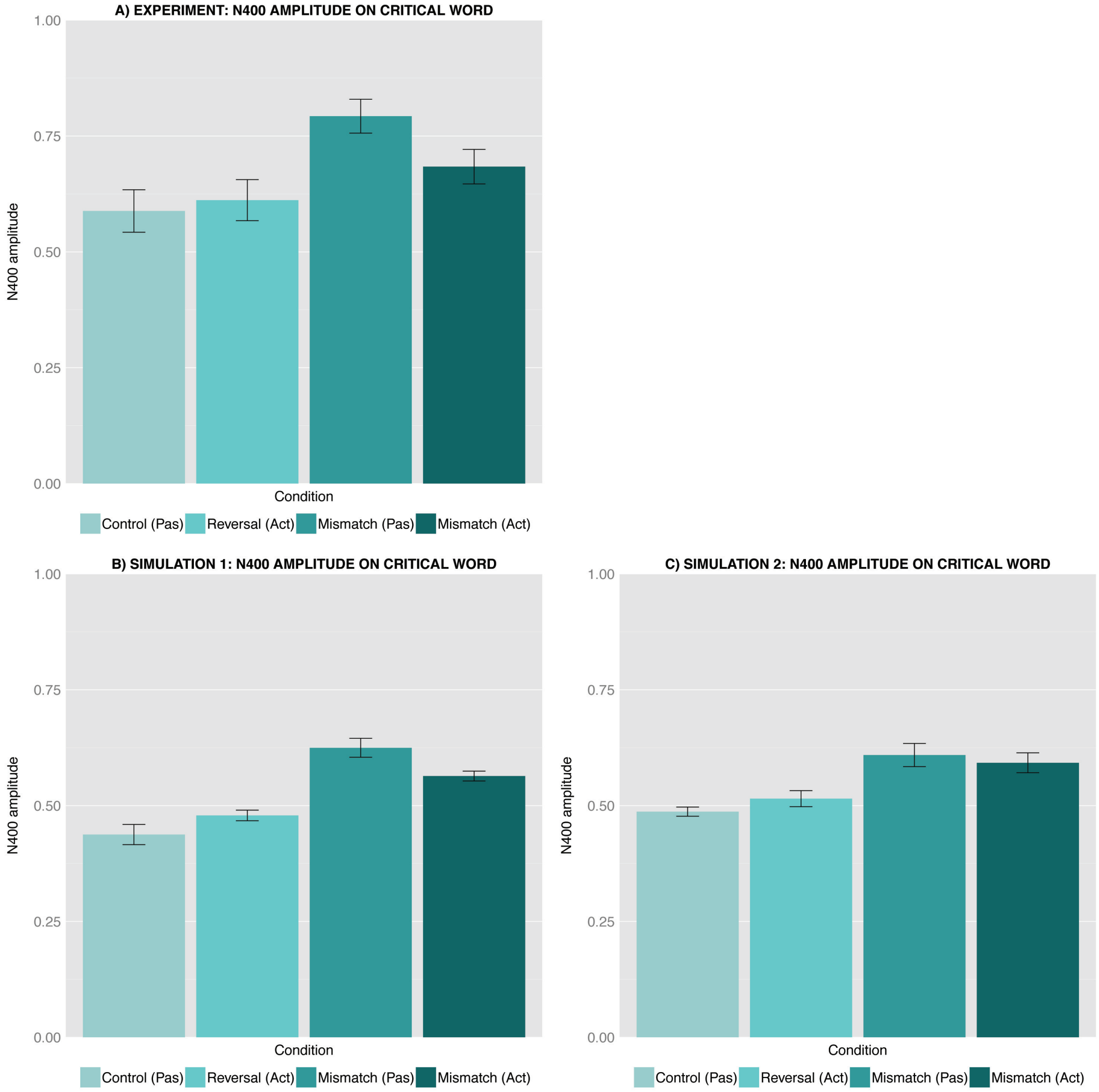

Fig. 4: N400 results. N400 results of the simulations in comparison to the results of the original experiment by Hoeks et al. (2004). Panel (a) shows the N400 amplitudes as measured in the original experiment (at the Pz electrode), transformed to a zero-to-one scale (see text). Panel (b) shows the N400 estimates measured in simulation 1, and panel (c) those measured in simulation 2. Error bars show standard errors.

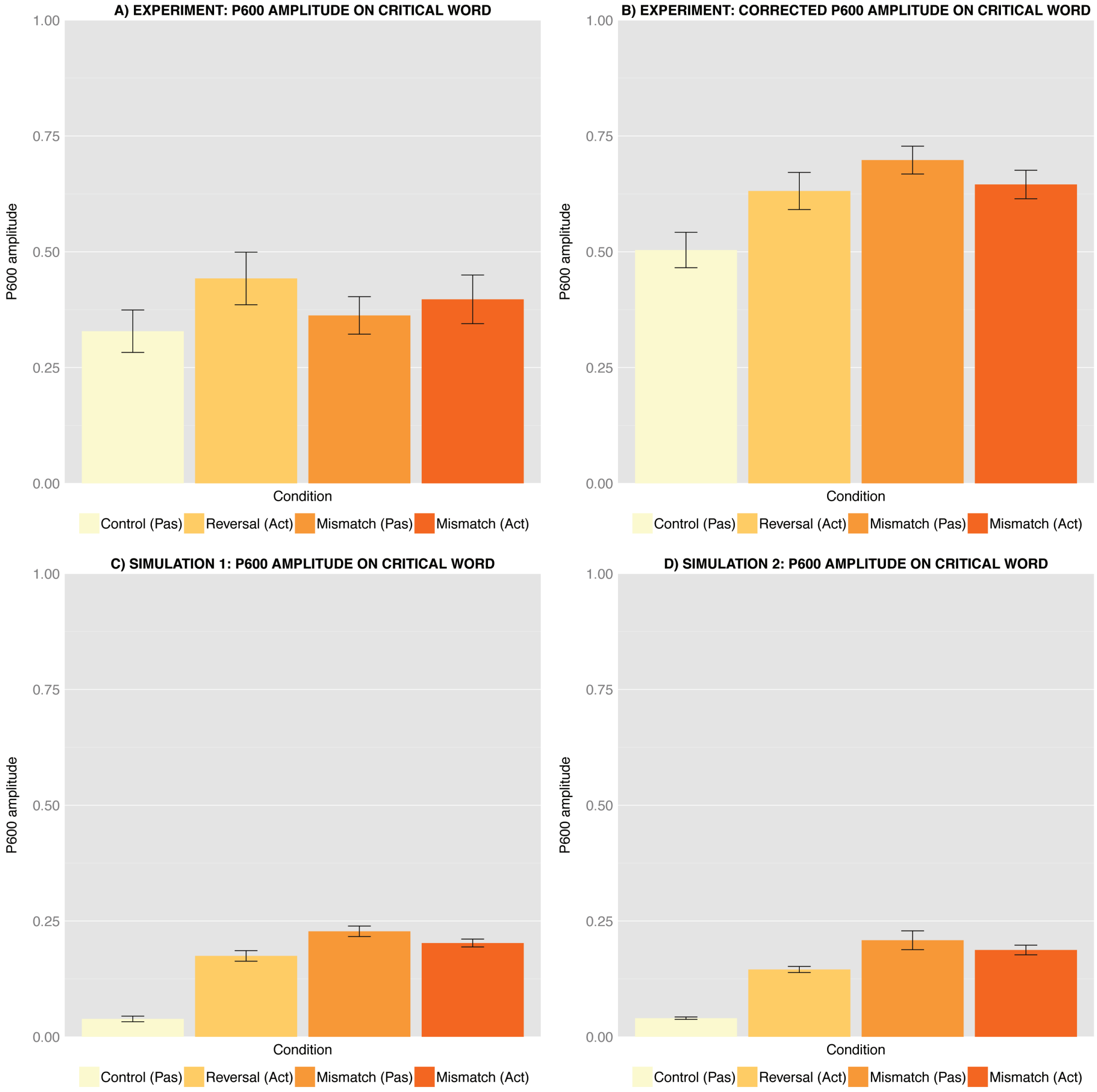

Fig. 5: P600 results. P600 results of the simulations in comparison to the results of the original experiment by Hoeks et al. (2004). Panel (a) shows the P600 amplitudes as measured in the original experiment (at the Pz electrode), transformed to a zero-to-one scale (see text). Panel (b) shows these P600 amplitudes corrected for overlap with the N400 component (also at Pz and on a zero-to-one scale) Panel (c) shows the P600 estimates measured in simulation 1, and panel (d) those measured in simulation 2. Error bars show standard errors.

4.4 On the ordering of effect sizes

The model successfully simulated the desired effects on all planned contrasts. An additional question is if the model also produces the same relative ordering of effects as revealed in the ERP data. For the N400, the model estimates do indeed numerically follow the ordering of the empirical data. However, for the P600, the model predicts P600 amplitude to be largest for the [Mismatch (Passive)] and [Mismatch (Active)] conditions, whereas in the empirical data it is largest for the [Reversal (Active)] condition. A possible reason for this difference is that in the ERP experiment the amplitude of the P600 may be affected by the amplitude of the preceding N400 (see Hagoort, 2003; Brouwer and Hoeks, 2013). If we correct for this, by pointwise subtraction of N400 amplitude from P600 amplitude in each condition, the pattern of results as far as the relative order within the three anomalous conditions is concerned comes in line with the results of the simulations (mismatch conditions larger than the role-reversed sentences; Fig. 5, top right). We take this to be an interesting observation for further study, and return to it in the discussion.

5 Discussion

We have presented a neurocomputational model that instantiates the recent Retrieval–Integration account of the N400 and the P600 in language processing (Brouwer et al., 2012; Brouwer and Hoeks, 2013). We have provided explicit and scalable (i.e., to larger groups neurons mimicking true cortical areas, as well as to models with larger empirical coverage) linking hypotheses between processing behavior in the model and estimates of the N400 and the P600, and we have shown that the model is able to capture traditional biphasic N400/P600-effects, as well as 'Semantic P600'-effects. Overall, we take our results to provide a 'proof of concept' of the RI account of the electrophysiology of semantic processing. Moreover, as our model instantiates a single-stream architecture, we take it to support the claim that there is no need for multiple processing streams (such as an independent semantic analysis stream) to explain 'Semantic P600'-effects. Below, we will discuss the implications of our model, and sketch directions for future research.

5.1 On the N400 and the role of context

On the Retrieval–Integration account, the retrieval of word meaning, reflected in N400 amplitude, is assumed to be contextually-driven. This poses an architectural constraint on the model, as we want its Retrieval module to instantiate this context-sensitivity. To this end, we trained the Retrieval module as part of the overall model, rather than as a separate module (see section 3.3.2). Our model successfully predicted the desired N400-effects for the contrasts tested in the Hoeks et al. (2004) study: no N400-effect for the role-reversed condition [Reversal (Active)] relative to control [Control (Passive)], and an N400-effect for both the passive mismatch ([Mismatch (Passive)] condition and the active mismatch condition [Mismatch (Active)] relative to control. In Appendix D, we show empirically that other approaches towards training the Retrieval module do not induce such context-sensitivity, and hence do not yield the desired results.

The reliance of the retrieval of word meaning on context supports the theoretical idea that the absence of an N400-effect for the role-reversed sentences 'De maaltijd heeft de kok bereid' (lit: 'The meal has the cook prepared') relative to controls 'De maaltijd werd door de kok bereid' (lit: 'The meal was by the cook prepared') is explained through contextual priming, stemming from both the preceding lexical items (e.g., 'meal' and 'cook'), as well as well as from the message representation that has been constructed so far (a scene involving a meal and a cook). Yet, a better understanding of where this and the other observed N400 patterns stem from in our model requires further scrutiny of the factors driving them. Each individual N400 estimate depends on two factors: the state of the RETRIEVAL layer after processing the pre-critical word (RETRIEVAL$_{t-1}$), and the state of the retrieval layer after processing the critical word (retrieval$_t$). Consequently, an N400-*effect* is governed by four factors: the two states of the retrieval layer in a target sentence (T:retrieval$_{t-1}$ and T:retrieval$_t$), and the two states of the retrieval layer in the control sentence (C:retrieval$_{t-1}$ and C:retrieval$_t$). More precisely, an N400-effect for a contrast may stem from differences between conditions at the pre-critical word (T:retrieval$_{t-1}$ $\neq$ C:retrieval$_{t-1}$), differences at the critical word (T:retrieval$_t$ $\neq$ C:retrieval$_t$), or both. To identify which of these factors govern our results, we numerically dissect the tested contrasts, using the sentences shown in Fig. 3 (from simulation 1): 'De maaltijd [werd door]/[heeft] de kok bereid/gezongen' (lit: 'The meal [was by]/[has] the cook prepared/sung).

In the [Mismatch (Passive)] versus [Control (Passive)] contrast, the sentences are identical up to the critical word ('De maaltijd werd door de kok'); that is, after processing the pre-critical word, there is no difference in the state of the retrieval layer between conditions (cos(T:retrieval$_{t-1}$,C:retrieval$_{t-1}$) $=$ 1). Hence, for this contrast the observed N400-effect is driven by the differences in the state of the retrieval layer induced by the critical word (cos(T:retrieval$_t$,C:retrieval$_t$) $=$ .268). In the [Mismatch (Active)] versus [Control (Passive)] contrast, however, the sentences differ prior to critical word, which is reflected in the differential state of the retrieval layer (cos(T:retrieval$_{t-1}$,C:retrieval$_{t-1}$) $=$ .879). At the critical word, however, the difference in the state of the retrieval layer is much more pronounced (cos(T:retrieval$_t$,C:retrieval$_t$) $=$ .235). Hence, the N400-effect for this contrast is again predominantly induced by the processing of the critical word. Indeed, the absence of an N400-effect for the [Reversal (Active)] versus [Control (Passive)] contrast lends further support for this view. Here, the difference at the pre-critical word is the same as in the [Mismatch (Active)] versus [Control (Passive)] contrast (cos(T:retrieval$_{t-1}$,C:retrieval$_{t-1}$) $=$ .879). Yet, at the critical word, the state of the retrieval layer is also highly similar across conditions (cos(T:retrieval$_t$,C:retrieval$_t$) $=$

.847). This tells us that, all things equal, a difference in voice (active/passive) only minimally affects the state of the RETRIEVAL layer. Hence, the N400-effects obtained in our simulations are primarily driven by the processing of the critical word. Of course, other manipulations of the pre-critical material, for instance, replacing one of the nouns, may have a stronger effect on the state of the RETRIEVAL layer at the pre-critical word, and hence on the N400 estimate at the critical word. In extending the model, this behavior may prove crucial for capturing N400-effects in context-manipulation designs.

Given that we estimate N400 amplitude as the dissimilarity of the RETRIEVAL layer at two consecutive time-steps, the model predicts an additional potential influence on the N400 amplitude: featural overlap between consecutive words. More specifically, our linking hypothesis predicts that N400 amplitude may be affected by the degree to which two consecutive words share semantic features in their meaning representations. This raises the question if such featural overlap is at play in our simulations, and if so, how they affect our N400-effects; that is, it could be that part of our N400-effects are driven by a larger similarity of the second noun (e.g., 'cook') to the congruent sentence-final verbs (e.g., 'prepared' in the [Control (Passive)] and [Reversal (Active)] conditions) compared to the incongruent sentence-final verbs (e.g., 'sung' in the [Mismatch (Passive)] and [Mismatch (Active)] conditions). The cosine similarities between the nouns and verbs, however, reveal that this was not the case. In simulation 1, there was only a small bias towards congruent continuations (congruence: .531 (se=.014); incongruence: .490 (.018)), whereas congruent and incongruent continuations were balanced in simulation 2 (congruence: .496 (.023); incongruence: .495 (.020)). If any at all, featural overlap between consecutive words should thus have only a very minimal effect on our N400-effects. Hence, our effects are being driven by context. However, this does not mean that featural overlap between consecutive words should generally not affect N400 amplitude; N400-effects in word pairs, for instance, in which a semantically unrelated second word of a pair produces a larger N400 than a semantically related one (Bentin et al., 1985; Boddy, 1981), might be largely driven by featural overlap, rather than contextual priming.

A further illustration of contextual modulation of our N400 estimates is the fact that the model correctly predicted the relative ordering of the N400-effects ([Control (Passive)] < [Reversal (Active)] < [Mismatch (Active)] < [Mismatch (Passive)]). Crucially, these relative differences can only be attributed to context, as the second nouns and final verbs are the same for the [Control (Passive)] and [Reversal (Active)] conditions ('kok/'cook' and 'bereid'/'prepared'), and for the [Mismatch (Passive)] and [Mismatch (Active)] conditions ('kok/'cook' and 'gezongen'/'sung').

5.2 Relation to other neurocomputational models

The simulations presented in the current paper focused on modeling the amplitudes of the N400 and the P600 component in sentence processing. Whereas our model is the first to capture the amplitude of both the N400 and the P600 component in a single neurocomputational model, our work is not the first attempt to model the processes underlying language-related ERP components. Recently, at least three other neurocomputational models have been put forward to explain certain aspects of the electrophysiology of language processing (all of which build on a rich history of neurocomputational—*connectionist*—models; see Christiansen and Chater, 2001 for an overview). These models differ in the type of linguistic processing that they aim to explain, as well as in the granularity of neurophysiological detail that they incorporate.

Crocker et al. (2010), for instance, propose a model of situated sentence processing that learns to mediate utterance and visual scene information, in order to construct a sentence interpretation in the form of a thematic role assignment representation (see also McClelland et al., 1989). This model produces P600 correlates for each word in a sentence, in a similar fashion as our model. However, the Crocker et al. model does not incorporate the N400, as we do in our model. The two other models focus on word recognition rather than sentence processing. Laszlo and Plaut (2012) propose a neurally plausible model of visual word recognition, which they show is able to successfully simulate N400 amplitude modulations during the reading of words, pseudowords, acronyms, and illegal strings, as well as to perform lexical decision. One aspect that is particularly noteworthy about this model is that it also successfully captures the temporal dynamics of the N400 component; that is the development of N400 amplitude over time. In the Laszlo and Plaut model, N400 amplitude is estimated as the mean semantic activation in the semantics layer of a connectionist architecture. Rabovsky and McRae (2014), however, challenge this relation between N400 amplitude and mean semantic activation. Using a set of simulations with a feature-based connectionist attractor network, they show that implicit prediction error—the difference between the semantic features that the model expected to encounter, and those actually encountered—provides a better account for N400 amplitude over a wide range of word processing phenomena, such as effects of semantic priming, semantic richness, frequency, and repetition. Interestingly, the idea of modeling N400 amplitude as implicit prediction error seems to be highly compatible with our approach towards modeling N400 amplitude as a measure of how much the pattern of activation changes in memory due to the processing of an incoming word; changes are large when pre-activated (implicitly predicted) features mismatch with the actual features of an incoming word. Although both the model by Laszlo and Plaut and the model by Rabovsky and McRae contribute to fine-grained insight into N400 modulations during word recognition, neither of the models captures N400-effects due to priming from the larger sentential context. Here, our model makes a novel contribution, as it is able to simulate contextually-induced N400 modulations. What is more, our model also produces estimates of sentence-level P600 modulations, which may prove difficult to incorporate in models of visual word recognition.

5.3 Towards covering a broader spectrum of ERP phenomena

In the present simulations, we focused on showing that our model can account for important patterns of ERP modulations in semantic processing. An important next step is to see if the model can also account for processing phenomena beyond semantically-induced effects, especially for the class of syntactically-induced ERP modulations, such as P600-effects to agreement violations (Hagoort et al., 1993) and garden-paths (Osterhout et al., 1994). On the RI account, these syntactically-induced P600-effects are taken to index difficulty in integrative processing, operating on the level of the utterance representation, rather than on the level of syntactic structure (see Brouwer et al., 2012, for further discussion). It is rather straightforward to see how this could explain the processing of garden-paths, as they entail a revision of the unfolding utterance interpretation. Indeed, here a 'syntactic structure'-based and an 'utterance representation'-based explanation are quite similar (i.e., revision of the analysis constructed thus far). The difference between these views, however, will become apparent if we turn to agreement violations, for which the RI account might at first glance seem less intuitive. A verb inducing an agreement violation, such as *throw* in 'The spoilt child <u>throw</u> [...]' has been shown to produce a P600-effect relative to a felicitous control 'The spoilt child <u>throws</u> [...]' (Hagoort et al., 1993). On a syntactic account of the P600, this reflects some kind of 'repair' of the infelicitous inflection on the verb (throw → throw*s*). On the RI account, however, we take this P600-effect to reflect difficulty in establishing a coherent utterance representation; the mismatch between *the spoilt child* and *throw* induces uncertainty about the input (cf. Levy, 2008), for instance about whether the speaker was talking about a single child or perhaps about multiple children (in which case not the inflection of the verb is incorrect, but the inflection of the noun).

To model the RI view on syntactically-induced ERP modulations, we need a richer representational scheme for utterance representations than thematic-role assignments. That is, we require a scheme that allows us, for instance, to differentiate between the different interpretations involved in the incremental processing of garden-path constructions, as well as to represent uncertainty about the singularity/plurality of agents and patients in agreement violations. In future work, we aim to replace the thematic-role assignment representations that serve as utterance representations in the current model, with richer utterance representations in terms of Distributed Situation Space (DSS) vectors (Frank et al., 2003, 2009).[4] Beyond capturing syntactically-induced P600 modulations, this approach will allow us to extend the model towards pragmatically-induced P600-effects (e.g., Hoeks et al., 2013; see Hoeks and Brouwer, 2014 for an overview).

5.4 Towards understanding processing in time

A general assumption in the literature is that the P600 component follows the N400 component in time. This entails that the processes generating the N400 component must be finished before those generating the P600 initiate. When applied to the Retrieval–Integration account, this means that the retrieval of word meaning must be fully completed, before integrative processing can commence. Our neurocomputational model of the RI account does indeed implement such a *serial* perspective: Retrieval of word meaning precedes Integrative processing. However, given the highly *parallel* nature of processing in the brain, this is almost certainly wrong. That is, the processes underlying the N400 and the P600 may actually overlap in time, and as ERPs are additive, such overlap between components means that P600 amplitude will be affected by N400 amplitude, and *vice versa* (see Hagoort, 2003; Kutas et al., 2006; Brouwer and Hoeks, 2013). This has important implications for the interpretation of ERP data (recall the discussion in section 4.4), which can be exemplified by means of a recent discussion on ERP effects in response to semantic anomalies.

Van Petten and Luka (2012) provide a review of 45 ERP studies in which semantically anomalous and non-anomalous sentence-final words are contrasted in sentences that

[4] For preliminary (unpublished) results using this approach, see Brouwer et al. (2015).

are otherwise (syntactically) correct. Out of a total of 64 comparisons, only 20 contrasts are reported to yield a significant Post-N400 Positivity (PNP). Upon careful examination of the critical waveforms of the reviewed studies (through digitization), however, we found that in only a small fraction of the comparisons the anomalous condition is numerically less positive than the non-anomalous condition. Hence, in the majority of contrasts, there is a numerical positivity for anomalous relative to the non-anomalous sentence-final words. This outcome is at odds with views that take the N400 to be a standalone deflection in the ERP signal. If pure semantic anomaly only affects N400 amplitude, without any influence on the P600, we should expect an approximately equal number of numerical post-N400 *positivities* and *negativities* (assuming the signal returns to pre-stimulus baseline). The prevalence of numerical positivities in the post-N400 time-window, then, suggests that the N400-effect for semantic anomaly is followed by a late positivity. A potential explanation for why these positivities fail to reach significance in the reviewed contrasts, is that P600 amplitude is reduced due to spatiotemporal overlap with the N400[5]. Indeed, in the majority of cases in which absence of a significant P600-effect is reported, the N400-effect is 'followed' by a numerical positivity, and in the few cases where this numerical positivity is not present, the N400-effect appears to be particularly large. If this kind of overlap is indeed at play, it not only applies to the review by van Petten and Luka (2012)—component overlap likely affects the majority of relevant ERP studies. This problem has been noted in the literature (Hagoort, 2003; Kutas et al., 2006; Brouwer and Hoeks, 2013), but it has remained unclear how to deal with it both theoretically in terms of (statistical) analysis of ERPs.

One approach towards getting to grips with component overlap, is to use explicit computational modeling to obtain a better understanding of the temporal dynamics of

[5] There is another factor that affects P600 amplitude: the processes underlying the P600 component are strongly task-dependent. That is, if an experiment does not involve any acceptability or plausibility judgment on a per-item basis, this typically leads to attenuated P600 amplitudes (Brouwer et al., 2012; see also Kolk et al., 2003). Critically, the studies reviewed by van Petten and Luka (2012) were explicitly selected to not involve any overt task, thereby biasing their results towards absence of P600-effects.

the processes indexed by the N400 and the P600. More specifically, instead of only estimating the word-elicited mean amplitude of these components, we propose to modify the neurocomputational model to estimate how the N400 and the P600 develop over time. To this end, a potential modification to the model will be to extend it with a more elaborate notion of time. That is, in the current model, word processing takes a single discrete time *tick*. Within this time tick, we take the retrieval of word meaning to precede its integration; that is, we take the feedforward propagation of activation from the INPUT to the INTEGRATION_OUTPUT layer to be itself a temporally extended process (activation first arrives at the RETRIEVAL layer, before it arrives at the RETRIEVAL_OUTPUT layer, and so forth). Consequently, N400 estimates temporally precede P600 estimates in the model, and hence do not overlap in time. In an extended model, word processing could be distributed over multiple time ticks (cf. Laszlo and Plaut, 2012), such that we obtain amplitude estimates of the N400 and the P600 at each of these. Component overlap can then be modeled, for instance, by combining the P600 estimate at time-step t with the N400 estimate at time-step $t + 1$. A first test-case for this extended model would be to see if it would get the ordering of the P600-effects in the Hoeks et al. (2004) study right, by assuming temporal overlap between the N400 and the P600.

5.5 Towards a cortical implementation of the model

At the core of our neurocomputational model are two relatively independent, but interacting sub-systems: a *retrieval* module and an *integration* module. Taken together, these sub-systems implement a high level language processing circuit. An important question is whether we can find out how this processing circuit might be neurally implemented in the brain. Brouwer and Hoeks (2013) provide a speculative answer to this question (see also Brouwer and Crocker, 2016). By focusing on computational *epicenters* (Mesulam, 1990, 1998) or *hubs* (Buckner et al., 2009), they propose a mapping of the RI account onto a minimal processing circuit. On this mapping, retrieval of word meaning

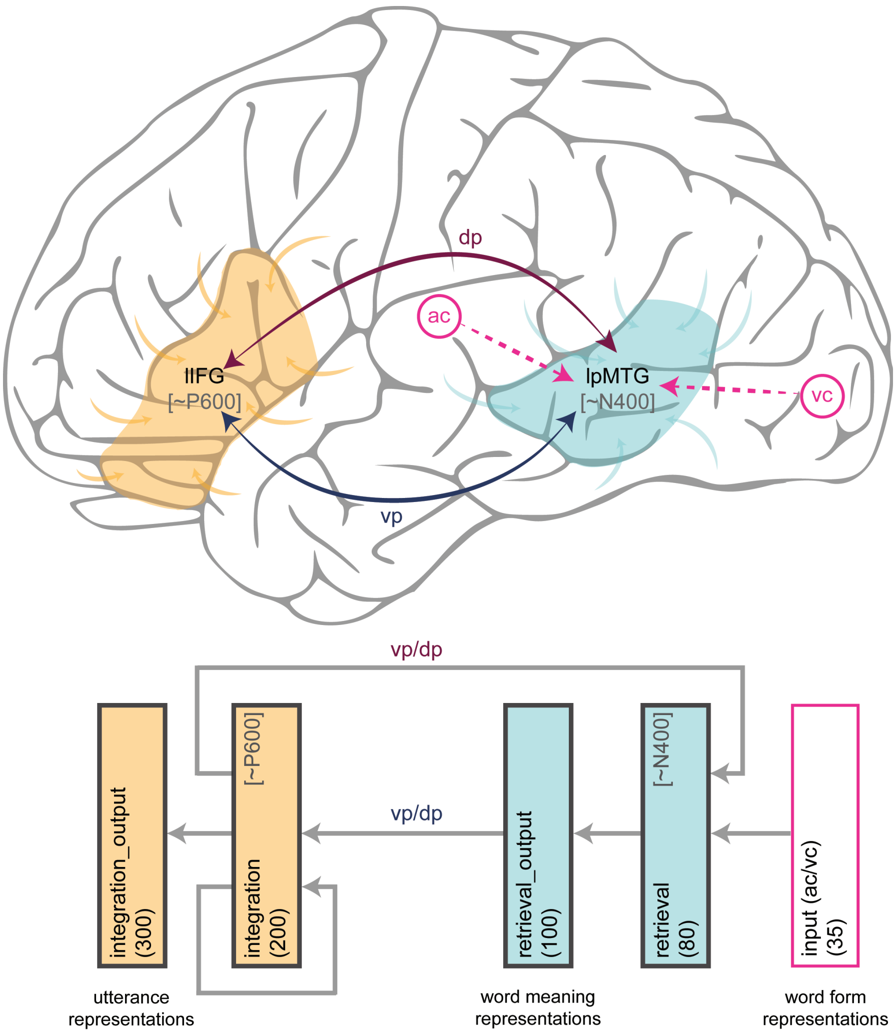

Fig. 6: A cortical implementation of the model. Schematic illustration of the functional-anatomic mapping of the RI account onto the core language network (top), and its relation to the neurocomputational model (bottom). Note that the schematic of the model uses a shorthand notation for the contextual input to the RETRIEVAL and INTEGRATION layers (by omitting the INTEGRATION_CONTEXT layer). An incoming word reaches the lpMTG via either the auditory cortex (ac) or visual cortex (vc) (corresponding to the INPUT layer in the model). The lpMTG then retrieves the conceptual knowledge associated with this word from the association cortices (RETRIEVAL → RETRIEVAL_OUTPUT), thereby generating the N400. Next, this retrieved meaning is sent to the lIFG (RETRIEVAL_OUTPUT → INTEGRATION), where it is integrated with its prior context (INTEGRATION → INTEGRATION) into an updated utterance representation (INTEGRATION → INTEGRATION_OUTPUT). This integration process is reflected in the P600 component. The updated utterance representation in the lIFG subsequently provides a context for the retrieval of the conceptual knowledge associated with the next word (INTEGRATION → RETRIEVAL).

is assumed to be mediated by the left posterior Middle Temporal Gyrus (lpMTG; BA 21), and this region is the presumed main generator underlying the N400 component. Integrative processing, on the other hand, is mediated by the left Inferior Frontal Gyrus (lIFG; BA 44/45/47), the presumed main generator of the P600 component. Brouwer and Hoeks (2013) provide a detailed derivation of this functional-anatomic mapping, which we will briefly summarize below.

On the basis of extenstive evidence from neuroimaging and lesion studies (Cabeza and Nyberg, 2000; Bookheimer, 2002; Dronkers et al., 2004; Lau et al., 2008; Binder et al., 2009; Price, 2010; Turken and Dronkers, 2011), Brouwer and Hoeks (2013) proposed the lpMTG (BA 21) as the epicenter mediating lexical retrieval (see Fig. 6, top). The recognition of spoken and written words, for instance, both in isolation and in sentences, consistently activates the lpMTG (Cabeza and Nyberg, 2000), and lesions in this region have been found to lead to difficulties in word-level comprehension (Dronkers et al., 2004). This suggests that the lpMTG is involved in word-form to meaning mapping (=lexical retrieval). What is more, the lpMTG is the only region that is active at both short and long stimulus onset asynchronies (SOAs) in lexical semantic priming studies (see Lau et al., 2008, for a large-scale review), supporting the idea that it plays a crucial role in the generation of the N400 component. This is consistent with source localization of the N400m, the magnetic field equivalent of the N400, which also points to the lpMTG as its main generator (Halgren et al., 2002). Finally, the lpMTG shows a particularly rich pattern of connectivity, both structurally and functionally, to frontal, parietal, and temporal regions of both hemispheres (Turken and Dronkers, 2011; see also Binder et al., 2009; Buckner et al., 2009; Koyama et al., 2010), supporting its role as a memory epicenter that retrieves and binds together conceptual knowledge from the association cortices, across which it is stored in a distributed manner.

Brouwer and Hoeks (2013) identified the lIFG (BA 44/45/47) as the epicenter mediating utterance (re)composition, and as such as the main generator of the P600 component (see Fig. 6, top). It has proven difficult to localize the neural generators underlying the P600 (Friederici, 2011). For one thing, consistent with the above hypothesis, a number of studies using fMRI have linked the P600 to the lIFG (see van de Meerendonk et al., 2011, for a discussion). However, attempts at reconstructing its generators using source localization have also identified the middle temporal gyrus and the posterior part of the temporal lobe as generator sites for the P600 (Kwon et al., 2005; Service et al., 2007). Brouwer and Hoeks (2013) tried to overcome these inconsistencies by looking for the generators of the P600 using a 'process alignment' strategy, where they attempt to align data regarding the nature and time-course of cognitive processing (from ERPs) with data on the cortical organization underlying it (from fMRI). This approach pointed them to the lIFG as the generator of the P600. The involvement of the lIFG in language processing is well-established, but the precise functional role of this region is a matter of ongoing debate (see Grodzinsky and Santi, 2008; Rogalsky and Hickok, 2011, for overviews). Rogalsky and Hickok (2011), for instance, point out that Broca's area (BA 44/45)—which is subsumed by the lIFG—has been hypothesized to be responsible for syntactic movement (Grodzinsky and Santi, 2008), hierarchical processing and phrase structure building (Friederici, 2009), order-related linearization processes (Bornkessel-Schlesewsky et al., 2009), working memory (Buchsbaum et al., 2005; Buchsbaum and D'Esposito, 2008), cognitive control (Novick et al., 2005), semantic unification (Hagoort, 2005), and thematic role checking and reanalysis (Caplan et al., 2008a,b). Others have stressed the role of the lIFG in the control of memory (Badre and Wagner, 2007). Rather than choosing between these different hypotheses, Brouwer and Hoeks (2013) proposed to unify them. That is, they proposed that the lIFG as a whole is host to a variety of processes involved in the (re)composition of an utterance representation, including syntax-based processes, semantic processes, (working) memory-related processes, and control processes, thereby subsuming the aforementioned hypotheses. Critically, recent anatomical investigations have identified a complex neuroarchitectural parcellation of the lIFG (Amunts et al., 2010; Amunts and Zilles, 2012), and Brouwer and Hoeks suggest that this parcellation may form the neuroanatomical basis for a fine-grained functional topology, in which different sub-processes of utterance representation (re)composition may be subserved by different, but potentially overlapping sub-parts of the lIFG (see Hagoort, 2005, Friederici, 2011, and Friederici and Singer, 2015 for similar ideas, also see Goucha and Friederici, 2015 for recent evidence supporting such a division of labor).

The word-associated conceptual knowledge retrieved by the lpMTG needs to be connected to the lIFG where it will be integrated into the current utterance representation, producing an updated utterance representation. Subsequently, this updated interpretation needs to be connected back to the lpMTG, such that contextual cues can pre-activate the conceptual knowledge that may be associated with possibly upcoming words. This information sharing requires bi-directional connectivity between the lpMTG and the lIFG. White matter fiber tractography using Diffusion Tensor Imaging (DTI) has led to the identification of a rather rich pattern of structural connectivity between the temporal and the frontal lobe (e.g., Catani et al., 2005; Saur et al., 2008; Makris and Pandya, 2009; Turken and Dronkers, 2011). This pattern consists of a dorsal pathway (dp) and a ventral pathway (vp) (see Fig. 6, top). The functional role of these pathways is, however, subject to an ongoing debate (Hickok and Poeppel, 2004, 2007; Saur et al., 2008; Friederici, 2009, 2011, 2012; Baggio and Hagoort, 2011; Tyler et al., 2011; Weiller et al., 2009; Bornkessel-Schlesewsky and Schlesewsky, 2013), which turns out to be difficult to settle as DTI does not allow for the determination of pathway directionality (Friederici, 2011). Nonetheless, the existence of these pathways clearly shows that there is white matter connectivity that supports the bi-directional information sharing between the lpMTG and lIFG required for RI processing cycles (see Brouwer and Hoeks, 2013, for further discussion).

Given the mapping of retrieval/N400 and integration/P600 onto respectively the lpMTG and the lIFG, and the existence of connectivity between these epicenters, Brouwer and Hoeks (2013) outlined a typical RI processing cycle. Depending on whether linguistic input is spoken or written, words reach the lpMTG via respectively the auditory cortex (ac) or the visual cortex (vc). The lpMTG then retrieves the conceptual knowledge associated with an incoming word from the association cortices. These retrieval processes generate the N400 component. The retrieved word meaning is then sent to the lIFG, via either the dorsal pathway (dp) or the ventral pathway (vp), where it is integrated with the current utterance representation, into an updated utterance representation. This integrative processing generates the P600 component. Finally, the updated utterance representation is connected back via one of the pathways (dp or vp) to the lpMTG, where it serves to pre-activate the conceptual knowledge associated with possible upcoming words. Fig. 6 (top) provides a schematic overview of the outlined cortical RI processing circuit. This circuit aligns well with the architecture of the neurocomputational model implementing the RI account (see Fig. 6), and hence provides a starting point for a more integrated functional-neuroanatomic model.

6 Conclusion

We have presented a neurocomputational model that instantiates the recent Retrieval–Integration account of the N400 and the P600 in semantic processing (Brouwer et al., 2012; Brouwer and Hoeks, 2013). We have provided explicit and scalable linking hypotheses of processing behavior in the model to estimates of the N400 and the P600, and we have shown that the model is able to capture both captures traditional biphasic N400/P600-effects and 'Semantic P600'-effects. Moreover, as our model instantiates a single-stream architecture, we take it to support the claim that there is no need for an independent semantic analysis stream to explain 'Semantic P600'-effects, and as such as evidence against multi-stream architectures. We also outlined how our model can be extended to expand the spectrum of ERP components that it accounts for and how it can be adapted to provide insights into the temporal dynamics of the processes underlying the N400 and the P600. Finally, we have speculated about a potential cortical instantiation of our model. We believe that by offering a formally precise neurocomputational implementation of the Retrieval–Integration account, we have not only elevated the debate on the functional interpretations of the N400 and the P600, but also raised the bar for competing models to move from conceptual to formally-precise descriptions of their assumed architecture, computational principles, and representations.

Acknowledgements

This research was funded by the Netherlands Organization for Scientific Research (NWO) PGW grant 10-26, awarded to the HB and JH. HB and MWC were also supported by the EU 7th Framework Programme Marie Curie Initial Training Network "Language and Perception" (LanPercept, grant Nr. 316748).

A Simulation materials

Table 2 lists the materials used in the simulations.

B Derivation of word meaning representations

As word meaning representations, our model employs 100-dimensional binary representations, which were derived from a large corpus of Dutch newspaper texts (the TwNC corpus; Ordelman et al., 2007) using the Correlated Occurrence Analogue to Lexical Semantics (COALS; Rohde et al., 2009).

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). This co-occurrence matrix, which we will refer to as X, is constructed for the 15.000 most frequent words. We then pruned all but the 14.000 columns of this matrix, 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:

$$w'_{a,b} = \frac{T \cdot w_{a,b} - \sum_j w_{a,j} \cdot \sum_i w_{i,b}}{(\sum_j w_{a,j} \cdot (T - \sum_j w_{a,j}) \cdot \sum_i w_{i,b} \cdot (T - \sum_i w_{i,b}))^{\frac{1}{2}}} \tag{5}$$

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

$$T = \sum_i \sum_j w_{i,j} \tag{6}$$

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

$$norm(w'_{a,b}) = \begin{cases} 0 & \text{if } w'_{a,b} < 0 \\ \sqrt{w'_{a,b}} & \text{otherwise} \end{cases} \tag{7}$$

To obtain the 100-dimensional feature vectors that we used in our simulations, we reduced the dimensionality of the normalized feature vectors by computing the Singular Value Decomposition of the co-occurrence matrix $X_{15000 \times 14000}$. 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:

$$\hat{X}_{15000 \times 14000} = \hat{U}_{15000 \times 100} \hat{S}_{100 \times 100} \hat{V}^T_{100 \times 14000} \tag{8}$$

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

$$V_c = X_c \hat{V} \hat{S}^{-1} \tag{9}$$

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

C Details of the training procedure

We trained each model (i.e., one for each simulation) using a two-stage training procedure (see sections 3.2 and 3.3). In both stages, the two models were trained using bounded gradient descent (Rohde, 2002), a modification of the standard backpropagation algorithm (Rumelhart et al., 1986). 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 INTEGRATION_OUTPUT layer:

$$E_c = \frac{1}{2}\sum_j (y_j - d_j)^2 \tag{10}$$

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:

$$\Delta w_{ij}(t) = -\varepsilon\rho\frac{\partial E}{\partial w_{ij}} + \alpha\Delta w_{ij}(t-1) \tag{11}$$

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

$$\rho = \begin{cases} \frac{1}{||\partial E/\partial w||} & \text{if } ||\partial E/\partial w|| > 1 \\ 1 & \text{otherwise} \end{cases} \tag{12}$$

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:

$$\frac{\partial E}{\partial w_{ij}} = \delta_j y_i \tag{13}$$

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

$$\delta_j = (y_j - d_j)(y_j(1 - y_j) + 0.1) \tag{14}$$

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

$$\delta_j = (y_j(1 - y_j) + 0.1) \sum_k \delta_k w_{jk} \tag{15}$$

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 items 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 an 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:

$$cos(x, y) = \frac{\sum_i x_i \times y_i}{\sqrt(\sum_i x_i^2) \times \sqrt(\sum_i y_i^2)} \tag{16}$$

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. For each of the models and after each training stage, comprehension performance was perfect (100% correct) on the training items. Finally, as the test items are a subset of the training items, comprehension performance was also perfect (100% correct) on the test sets.

D Training on perfect word meaning representations

The Retrieval module of our model was trained using a rather non-standard training procedure; we trained it as part of the overall network, rather than as a separate network (see section 3.3.2 for details). We argued that this training procedure is necessary to pressure the model to arrive at a context-sensitive solution in the Retrieval module. Here, we compare the results of this training regime to those obtained with a training procedure in which the Retrieval module is trained on correct word meaning representations (COALS vectors) at the RETRIEVAL_OUTPUT layer (see Table 3). More specifically, we compare the results of our model to four new models, which differ in various architectural aspects. Each of these models is derived by taking the trained Integration module from our model, and then training the Retrieval module on word meaning representations using the same procedure and parameters as discussed above (with the exception that training only lasted 700 epochs, as the models converged faster).

Two of these models have architectures identical to our neurocomputational model (TRUEMODEL), but their Retrieval modules were trained on perfect word meaning representations: the **IntegrationContext** model and the **PerfectIntegrationContext** model. In the **IntegrationContext** model, the contexts in the INTEGRATION_CONTEXT layer depend on the quality of the word meaning representations produced at the RETRIEVAL_OUTPUT layer during training, whereas in the **PerfectIntegrationContext** model these contexts were perfect (i.e., they were recorded from the Integration module). A first thing to note is that both models produce the same P600-effects as our neurocomputational model, which is due the fact that the Integration module is unchanged; only its inputs differ slightly. Neither of them, however, produces the desired pattern of N400-effects; differences between conditions are minimal, and the ordering of N400 estimates is wrong. In a third model, the **RetrievalContext** model, the Retrieval module is trained as a separate SRN with only its own local context (i.e., a RETRIEVAL_CONTEXT layer which receives a copy from the RETRIEVAL layer prior to feedforward propagation, and a RETRIEVAL_CONTEXT → RETRIEVAL projection). Again, whereas this model produces the same P600-effects as our model, it fails to produce the desired N400-effects (minimal differences and incorrect ordering). Finally, the **NoContext** model, is a model in which the RETRIEVAL layer receives no contextual information at all. This model also produces the P600-effects our model produces, but not the N400-effects (again, minimal differences and incorrect ordering).

References

Amunts, K., Lenzen, M., Friederici, A. D., Schleicher, A., Morosan, P., Palomero-Gallagher, N., and Zilles, K. (2010). Broca's region: Novel organizational principles and multiple receptor mapping. *PLoS Biology*, 8(9).

Amunts, K. and Zilles, K. (2012). Architecture and organizational principles of broca's region. *Trends in Cognitive Sciences*, 16(8):418–426.

Badre, D. and Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia*, 45(13):2883–2901.

Baggio, G. and Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, 26:1338–1367.

Bentin, S., McCarthy, G., and Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, 60(4):343–355.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796.

Boddy, J. (1981). Evoked potentials and the dynamics of language processing. *Biological Psychology*, 13:125–140.

Bookheimer, S. (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience*, 25(1):151–188.

Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2008). An alternative perspective on semantic P600 effects in language comprehension. *Brain Research Reviews*, 59(1):55–73.

Bornkessel-Schlesewsky, I. and Schlesewsky, M. (2013). Reconciling time, space and function: A new dorsal–ventral stream model of sentence comprehension. *Brain and language*, 125(1):60–76.

Bornkessel-Schlesewsky, I., Schlesewsky, M., and von Cramon, D. Y. (2009). Word order and broca's region: Evidence for a supra-syntactic perspective. *Brain and Language*, 111(3):125–139.

Brouwer, H. and Crocker, M. W. (2016). On the organization of the perisylvian cortex: Insights from the electrophysiology of language: Comment on "towards a computational comparative neuroprimatology: Framing the language-ready brain" by M.A. Arbib. *Physics of Life Reviews*, 16:58–60.

Brouwer, H., Fitz, H., and Hoeks, J. C. J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, 1446:127–143.

Brouwer, H. and Hoeks, J. C. J. (2013). A time and place for language comprehension: Mapping the N400 and the P600 to a minimal cortical network. *Frontiers in Human Neuroscience*, 7:758.

Brouwer, H., Hoeks, J. C. J., and Crocker, M. W. (2015). The electrophysiology of language comprehension: A neurocomputational model. In *Society for the Neurobiology of Language (SNL2015)*.

Buchsbaum, B. R. and D'Esposito, M. (2008). The search for the phonological store: From loop to convolution. *Journal of Cognitive Neuroscience*, 20(5):762–778.

Buchsbaum, B. R., Olsen, R. K., Koch, P., and Berman, K. F. (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron*, 48(4):687–697.

Buckner, R. L., Sepulcre, J., Talukdar, T., Krienen, F. M., Liu, H., Hedden, T., Andrews-Hanna, J. R., Sperling, R. A., and Johnson, K. A. (2009). Cortical hubs revealed by intrinsic functional connectivity: Mapping, assessment of stability, and relation to Alzheimer's disease. *The Journal of Neuroscience*, 29(6):1860–1873.

Cabeza, R. and Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience*, 12(1):1–47.

Caplan, D., Chen, E., and Waters, G. (2008a). Task-dependent and task-independent neurovascular responses to syntactic processing. *Cortex*, 44(3):257–275.

Caplan, D., Stanczak, L., and Waters, G. (2008b). Syntactic and thematic constraint effects on blood oxygenation level dependent signal correlates of comprehension of relative clauses. *Journal of Cognitive Neuroscience*, 20(4):643–656.

Catani, M., Jones, D. K., and ffytche, D. H. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, 57(1):8–16.

Christiansen, M. H. and Chater, N. (2001). *Connectionist psycholinguistics*. Greenwood Publishing Group.

Crocker, M. W., Knoeferle, P., and Mayberry, M. R. (2010). Situated sentence processing: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, 112(3):189–201.

Dronkers, N. F., Wilkins, D. P., van Valin, R. D., Redfern, B. B., and Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1):145–177.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Fahlman, S. E. (1988). An empirical study of learning speed in back-propagation networks. Technical report, Carnegie Mellon University.

Frank, S. L., Haselager, W. F. G., and van Rooij, I. (2009). Connectionist semantic systematicity. *Cognition*, 110(3):358–379.

Frank, S. L., Koppen, M., Noordman, L. G. M., and Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, 27(6):875–910.

Friederici, A. D. (2009). Pathways to language: Fiber tracts in the human brain. *Trends in Cognitive Sciences*, 13(4):175–181.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4):1357–1392.

Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, 16(5):262–268.

Friederici, A. D. and Singer, W. (2015). Grounding language processing on basic neurophysiological principles. *Trends in Cognitive Sciences*.

Goucha, T. and Friederici, A. D. (2015). The language skeleton after dissecting meaning: A functional segregation within brocaÂŠs area. *NeuroImage*.

Grodzinsky, Y. and Santi, A. (2008). The battle for broca's region. *Trends in Cognitive Sciences*, 12(12):474–480.

Hagoort, P. (2003). Interplay between syntax and semantics during sentence comprehension: ERP effects of combining syntactic and semantic violations. *Journal of Cognitive Neuroscience*, 15(6):883–899.

Hagoort, P. (2005). On broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, 9(9):416–423.

Hagoort, P., Baggio, G., and Willems, R. M. (2009). Semantic unification. In Gazzaniga, M. S., editor, *The Cognitive Neurosciences, 4th ed.*, pages 819–836. MIT Press.

Hagoort, P., Brown, C., and Groothusen, J. (1993). The Syntactic Positive Shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, 8(4):439–483.

Halgren, E., Dhond, R. P., Christensen, N., van Petten, C., Marinkovic, K., Lewine, J. D., and Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage*, 17(3):1101–1116.

Hickok, G. and Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1):67–99.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Hoeks, J. C. J. and Brouwer, H. (2014). Electrophysiological research on conversation and discourse processing. In Holtgraves, T. M., editor, *The Oxford Handbook of Language and Social Psychology*, pages 365–386. New York: Oxford University Press.

Hoeks, J. C. J., Stowe, L. A., and Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19(1):59–73.

Hoeks, J. C. J., Stowe, L. A., Hendriks, P., and Brouwer, H. (2013). Questions left unanswered: How the brain responds to missing information. *PLoS ONE*, 8(10).

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

Kolk, H. H. J., Chwilla, D. J., van Herten, M., and Oor, P. J. W. (2003). Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, 85(1):1–36.

Koyama, M. S., Kelly, C., Shehzad, Z., Penesetti, D., Castellanos, F. X., and Milham, M. P. (2010). Reading networks at rest. *Cerebral Cortex*, 20(11):2549–2559.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49.

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

Kutas, M. and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12):463–470.

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

Kutas, M., van Petten, C., and Kluender, R. (2006). Psycholinguistics electrified II: 1994–2005. In Traxler, M. J. and Gernsbacher, M. A., editors, *Handbook of Psycholinguistics, 2nd Edition*, pages 659–724. Elsevier, New York.

Kwon, H., Kuriki, S., Kim, J. M., Lee, Y. H., Kim, K., and Nam, K. (2005). MEG study on neural activities associated with syntactic and semantic violations in spoken korean sentences. *Neuroscience Research*, 51(4):349–357.

Laszlo, S. and Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, 120(3):271–281.

Lau, E. F., Phillips, C., and Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, 9(12):920–933.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing*, pages 234–243. Association for Computational Linguistics.

Makris, N. and Pandya, D. N. (2009). The extreme capsule in humans and rethinking of the language circuitry. *Brain Structure and Function*, 213(3):343–358.

Mayberry, M. R., Crocker, M. W., and Knoeferle, P. (2009). Learning to attend: A connectionist model of situated language comprehension. *Cognitive Science*, 33(3):449–496.

McClelland, J. L., St. John, M. F., and Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4(3-4):SI287–SI335.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Mesulam, M. M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology*, 28(5):597–613.

Mesulam, M. M. (1998). From sensation to cognition. *Brain*, 121(6):1013.

Nieuwland, M. S. and van Berkum, J. J. A. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3):691–701.

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, 5(3):263.

Ordelman, R., Jong, F., Hessen, A., and Hondorp, H. (2007). TwNC: A multifaceted dutch news corpus. *ELRA Newsletter*, 12(3-4).

Osterhout, L., Holcomb, P. J., and Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4):786–803.

Price, C. J. (2010). The anatomy of language: A review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, 1191(1):62–88.

Rabovsky, M. and 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.

Rogalsky, C. and Hickok, G. (2011). The role of Broca's area in sentence comprehension. *Journal of Cognitive Neuroscience*, 23(7):1664–1680.

Rohde, D. L. T. (2002). *A connectionist model of sentence comprehension and production*. PhD thesis, Carnegie Mellon University.

Rohde, D. L. T., Gonnerman, L. M., and Plaut, D. C. (2009). An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*, pages 1–33.

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

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M., Umarova, R., Musso, M., Glauche, V., Abel, S., Huber, S., Rijntjes, M., Hennig, J., and Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences*, 105(46):18035–18040.

Service, E., Helenius, P., Maury, S., and Salmelin, R. (2007). Localization of syntactic and semantic brain responses using magnetoencephalography. *Journal of Cognitive Neuroscience*, 19(7):1193–1205.

Turken, A. U. and Dronkers, N. F. (2011). The neural architecture of the language comprehension network: Converging evidence from lesion and connectivity analyses. *Frontiers in Systems Neuroscience*, 5:1.

Tyler, L. K., Marslen-Wilson, W. D., Randall, B., Wright, P., Devereux, B., Zhuang, J., Papoutsi, M., and Stamatakis, E. A. (2011). Left inferior frontal cortex and syntax: Function, structure and behaviour in patients with left hemisphere damage. *Brain*, 134(2):415–431.

van Berkum, J. J. A. (2009). The ‘neuropragmatics’ of simple utterance comprehension: An ERP review. In Sauerland, U. and Yatsushiro, K., editors, *Semantics and Pragmatics: From experiment to theory*, pages 276–316. Palgrave Macmillan, Basingstoke.

van de Meerendonk, N., Indefrey, P., Chwilla, D. J., and Kolk, H. H. J. (2011). Monitoring in language perception: Electrophysiological and hemodynamic responses to spelling violations. *NeuroImage*, 54(3):2350–2363.

van Petten, C. and Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International Journal of Psychophysiology*, 83(2):176–190.

Weiller, C., Musso, M., Rijntjes, M., and Saur, D. (2009). Please don’t underestimate the ventral pathway in language. *Trends in Cognitive Sciences*, 13(9):369–1370.

Table 2: Simulation materials. Overview of the materials used in the simulations. The upper part of this table represents the lexical items used in simulation 1, and the bottom half those in simulation 2.

Sim.	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*
Sim.	**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 3: Comparison of various training regimes for the Retrieval module. Mean N400 and P600 estimates (and standard errors in parentheses) for our neurocomputational model (TrueModel), compared to four different models trained on perfect word meaning representations (COALS vectors). CP = Control (Passive); RA = Reversal (Active); MP = Mismatch (Passive); MA = Mismatch (Active). See text for details on the models.

Model	**Condition**	**Simulation 1**		**Simulation 2**	
		N400	**P600**	**N400**	**P600**
TrueModel	CP	**.438** (.022)	.039 (.006)	**.487** (.010)	.040 (.003)
	RA	**.479** (.011)	.175 (.011)	**.515** (.017)	.145 (.007)
	MP	**.625** (.020)	.228 (.011)	**.609** (.025)	.208 (.020)
	MA	**.564** (.011)	.202 (.009)	**.592** (.021)	.187 (.010)
IntegrationContext	CP	**.355** (.007)	.066 (.007)	**.355** (.006)	.064 (.005)
	RA	**.349** (.008)	.200 (.010)	**.355** (.006)	.165 (.012)
	MP	**.366** (.010)	.230 (.010)	**.352** (.010)	.212 (.020)
	MA	**.368** (.012)	.216 (.009)	**.357** (.010)	.203 (.012)
PerfectIntegrationContext	CP	**.346** (.007)	.066 (.007)	**.351** (.007)	.063 (.005)
	RA	**.340** (.009)	.198 (.010)	**.351** (.007)	.166 (.012)
	MP	**.364** (.009)	.230 (.009)	**.354** (.010)	.214 (.021)
	MA	**.365** (.010)	.216 (.009)	**.362** (.010)	.204 (.012)
RetrievalContext	CP	**.330** (.006)	.064 (.007)	**.356** (.010)	.060 (.005)
	RA	**.333** (.005)	.196 (.010)	**.353** (.010)	.159 (.011)
	MP	**.345** (.007)	.223 (.010)	**.356** (.009)	.207 (.020)
	MA	**.347** (.007)	.208 (.009)	**.353** (.009)	.197 (.012)
NoContext	CP	**.297** (.004)	.064 (.007)	**.315** (.008)	.062 (.005)
	RA	**.297** (.004)	.196 (.010)	**.315** (.008)	.164 (.012)
	MP	**.315** (.007)	.229 (.010)	**.307** (.007)	.211 (.020)
	MA	**.315** (.007)	.213 (.009)	**.307** (.007)	.201 (.011)