N400 amplitudes as change in a probabilistic representation of meaning: A neural network model

Milena Rabovsky*, Steven S. Hansen, & James L. McClelland*

Department of Psychology, Stanford University

Word count: 4064

*Corresponding authors:

Milena Rabovsky (milena.rabovsky@gmail.com)

James L. McClelland (mcclelland@stanford.edu)

Abstract

The N400 component of the event-related brain potential has aroused much interest because it is thought to provide an online measure of meaning processing in the brain. Yet, the underlying process of meaning construction remains incompletely understood. Here, we present a computationally explicit account of this process and the emerging representation of sentence meaning. We simulate N400 amplitudes as the change induced by an incoming stimulus in an implicit and probabilistic representation of meaning captured by the hidden unit activation pattern in a neural network model of sentence comprehension, and propose that the process underlying the N400 also drives implicit learning in the network. We account for a broad range of empirically observed N400 effects which have previously been difficult to capture within a single integrated framework.

N400 amplitudes as change in a probabilistic representation of meaning:

A neural network model

The N400 component of the event-related brain potential (ERP) has received a great deal of attention, as it promises to shed light on the brain processes underlying the semantic interpretation of language and other meaningful inputs. The N400 is a negative deflection at centroparietal electrode sites peaking around 400 ms after the presentation of a potentially meaningful stimulus. The seminal N400 study showed that N400 amplitude varies as a function of meaning in context: given "I take my coffee with cream and ..." the anomalous word *dog* produces a larger N400 than the congruent word *sugar*.[1] Since this study, the N400 has been used as a dependent variable in over 1000 studies and has been shown to be modulated by a wide variety of variables including sentence context, categorical relations, repetition, and lexical frequency, amongst others[2]. However, despite the large amount of data on the N400, its functional basis is not well understood: various verbal descriptive theories have been actively debated, proposing, for instance, that N400 amplitudes reflect lexical and/or semantic access[3], semantic integration[4,5], semantic binding[6], or semantic inhibition[7].

Here, we provide both support for and formalization of the view that the N400 reflects the stimulus-driven update of a representation of sentence meaning – one that implicitly and probabilistically represents all aspects of meaning as it evolves in real time during comprehension[2]. We do so by presenting an explicit computational model of this process, showing that it can account for a broad range of empirically observed N400 effects which have been difficult to capture within a single theoretical account[2]. Rather than concentrating on neurophysiological details, we directly relate variations in N400 amplitudes to measures obtained from a more abstract, functional level account, allowing us to focus on the goal of clarifying the cognitive functions underlying N400 amplitudes.

The design of the model[8] reflects the principle that listeners continually update an implicit probabilistic representation of sentence meaning as each incoming word of a sentence is presented. The representation is an internal representation (corresponding to a pattern of neural activity, modeled in an artificial neural network) called the *sentence gestalt* (SG) that depends on connection-based knowledge in the *update* part of the network (see Fig. 1). The SG pattern can be characterized as implicitly representing subjective probability distributions over the various possible aspects or attributes of the event being described by the sentence (see *Implicit probabilistic theory of sentence meaning* section in *online methods*). The magnitude of the update produced by each successive word corresponds to the change in this implicit representation that is produced by the word, and it is this change, we propose, that is reflected in N400 amplitudes. For example, after a listener has heard "I take my coffee with cream and…" our account holds that the activation state already implicitly represents the belief that the speaker takes her coffee with cream and sugar, so the representation will change very little when the final word "…sugar" is presented, resulting in little or no N400 signal; in contrast, the representation will change much more if "…dog" is presented instead, corresponding to a much larger change in belief and a larger N400. Specifically, the *semantic update* (SU) induced by the current word *n* is defined as the sum across the SG layer units of the absolute value of the change in the unit's activation produced by the current word *n*, i.e. the difference in the unit's activation after word *n* and after word *n-1*:

$$N400_n = SU_n = \sum_i |a_i(w_n) - a_i(w_{n-1})|$$

This measure can be related formally to a Bayesian measure of surprise[9] and to the signals that govern learning in the network (see *online methods* and below), and indeed we propose a new learning rule driven by the semantic update.

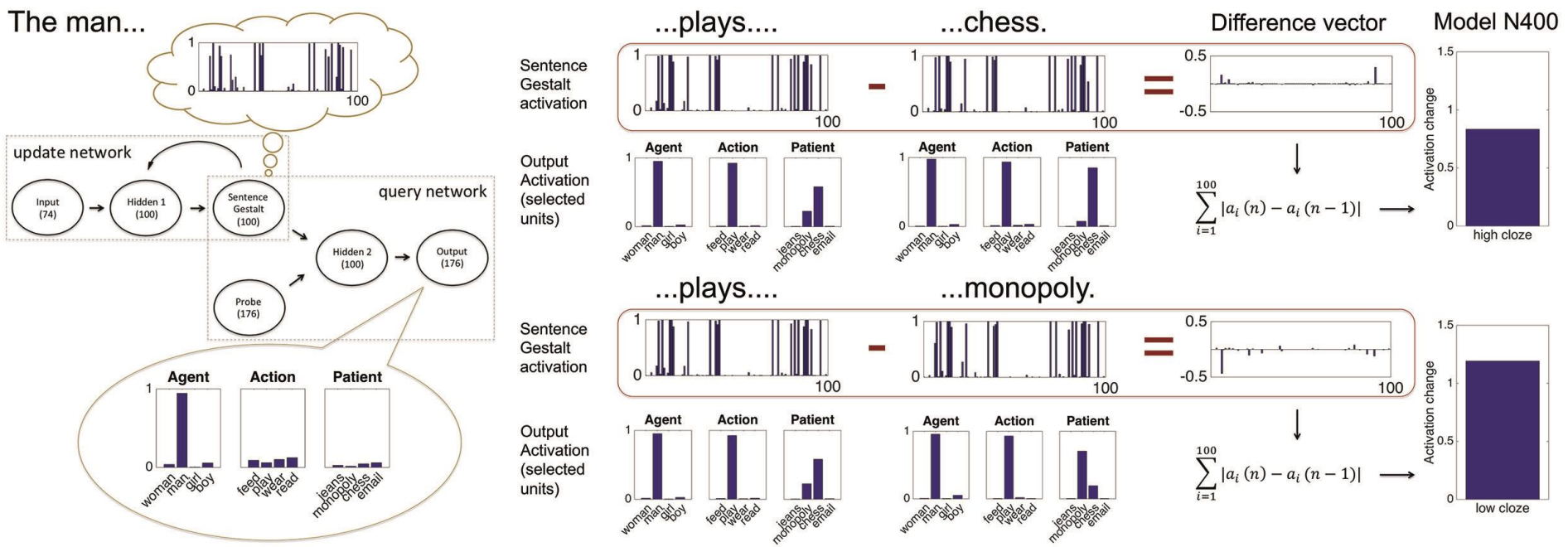


Figure 1. *The Sentence Gestalt (SG) model architecture (left), processing sentences with a high (top) or low (bottom) cloze probability ending, and computation of N400 correlate (right). Grey boxes highlight the model's parts. Ovals represent layers of units (numbers give the number of units in each layer). Arrows represent all-to-all modifiable connections; each unit applies a sigmoid transformation to its summed inputs, where each input is the product of the activation of the sending unit times the weight of that connection. In the update part of the model, sequentially incoming words are processed through a first hidden layer (Hidden 1) where their input combines with the previous activation of the SG layer to produce the updated SG layer activation (shown as a vector above the model) corresponding to the model's current probabilistic representation of the meaning of the sentence. During training, after each presented word, the model is probed concerning all aspects of the described event (e.g. agent, "man", action, "play", patient, "monopoly", etc.) in the query part of the network. Here, the activation from the probe layer combines via hidden layer 2 with the current SG activation to produce output activations (output units for selected specific concepts activated in response to the agent, action, and patient probes are shown; each concept is also represented by semantic feature units; see Supplementary Table 1). After presentation of "The man" (leftmost), the model activates the correct unit when probed for the agent, and estimates the probabilities of the action and patient of the event. When presented with the second word "plays" the SG activation is updated (second 100 unit vector) and the model now activates the correct output units to the agent and action probe, and updates its activations estimating the probability of each possible patient. These estimates reflect the model's experience, since chess occurs with relatively high probability. Thus, if the next word is "chess", less update of the SG layer activations is necessary (top) than if the next word in "monopoly" (bottom). The computation of the model's N400 correlate (right), called the Semantic Update (SU), reflects the change in the SG activation induced by the current word; the SU is larger for low (monopoly, bottom) as compared to high cloze probability (chess, top) endings.*

Results

The representations that a listener forms when encountering the words in a sentence reflect the statistics of past experience with events and the sentences that describe them and are thought to depend on exposure to sentences in contexts where information about the events the sentences describe is also available (see Fig. 1 and *online methods*). To test the model, we use an artificial corpus of {*sentence*, *event*} training examples produced by a generative model that embodies specific assumptions about the statistics of events and sentences (see *online methods*). While the sentences and event descriptions are simpler than real sentences and events, the artificial corpus has the advantage that its properties can be completely understood, allowing separate manipulation of statistical and semantic relationships, which are difficult to fully separate in natural corpora. We report twelve simulations of well-established N400 effects chosen to illustrate how the model can address empirical findings taken as supporting diverse and sometimes conflicting descriptive theories of the functional basis of the N400 signal (see Table 1). We focus on language-related effects but note that both linguistic and non-linguistic information contribute to changes in semantic activation as reflected by the N400[2].

Please insert Table 1 about here

Basic effects

From "violation signal" to graded reflection of surprise. The N400 was first observed after a semantically anomalous sentence completion such as e.g. "He spread the warm bread with *socks*"[1] as compared to a high probability congruent completion (*butter*). Correspondingly, in our model, SU was significantly larger for sentences with endings that are both semantically and statistically inconsistent with the training corpus compared to semantically consistent, high-probability completions (Fig. 2a and Supplementary Fig. 1a).

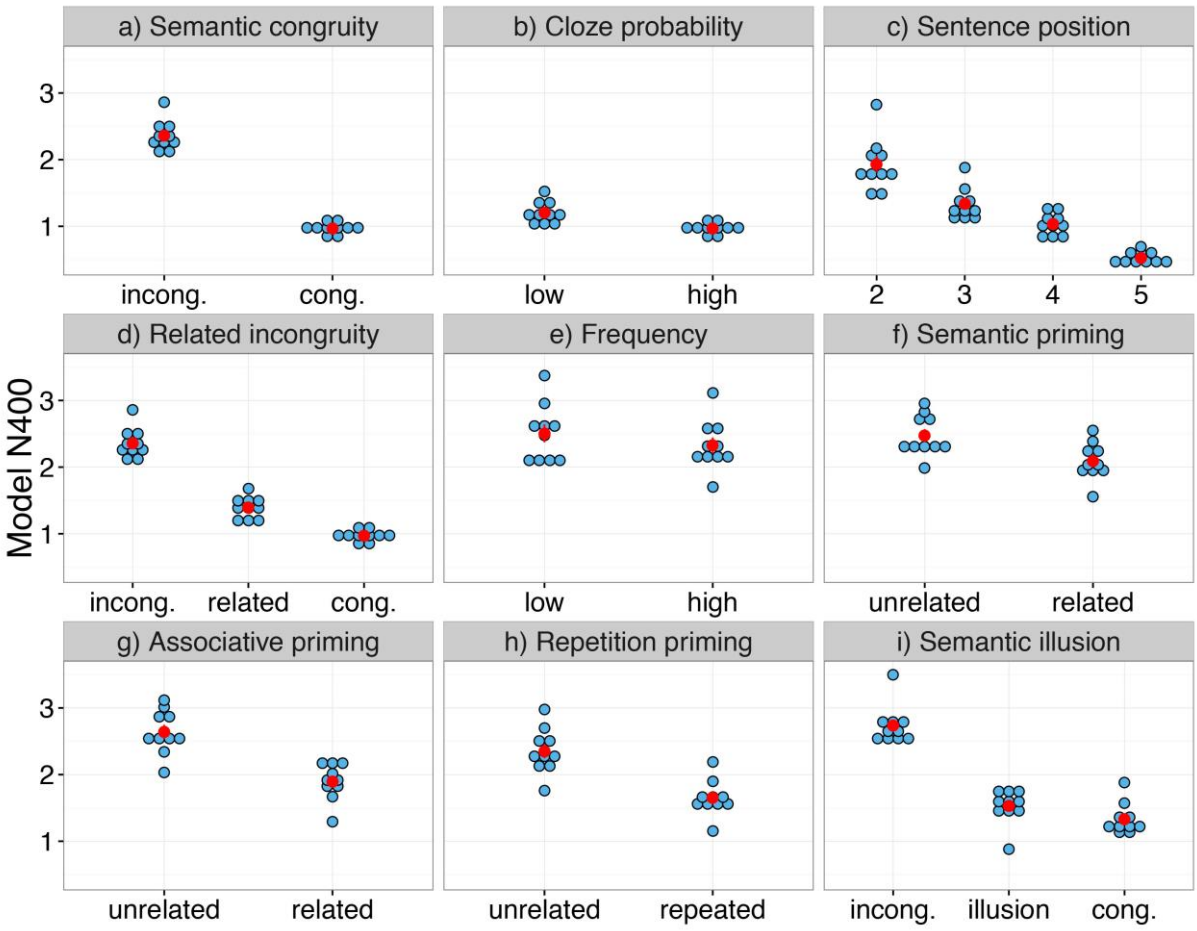
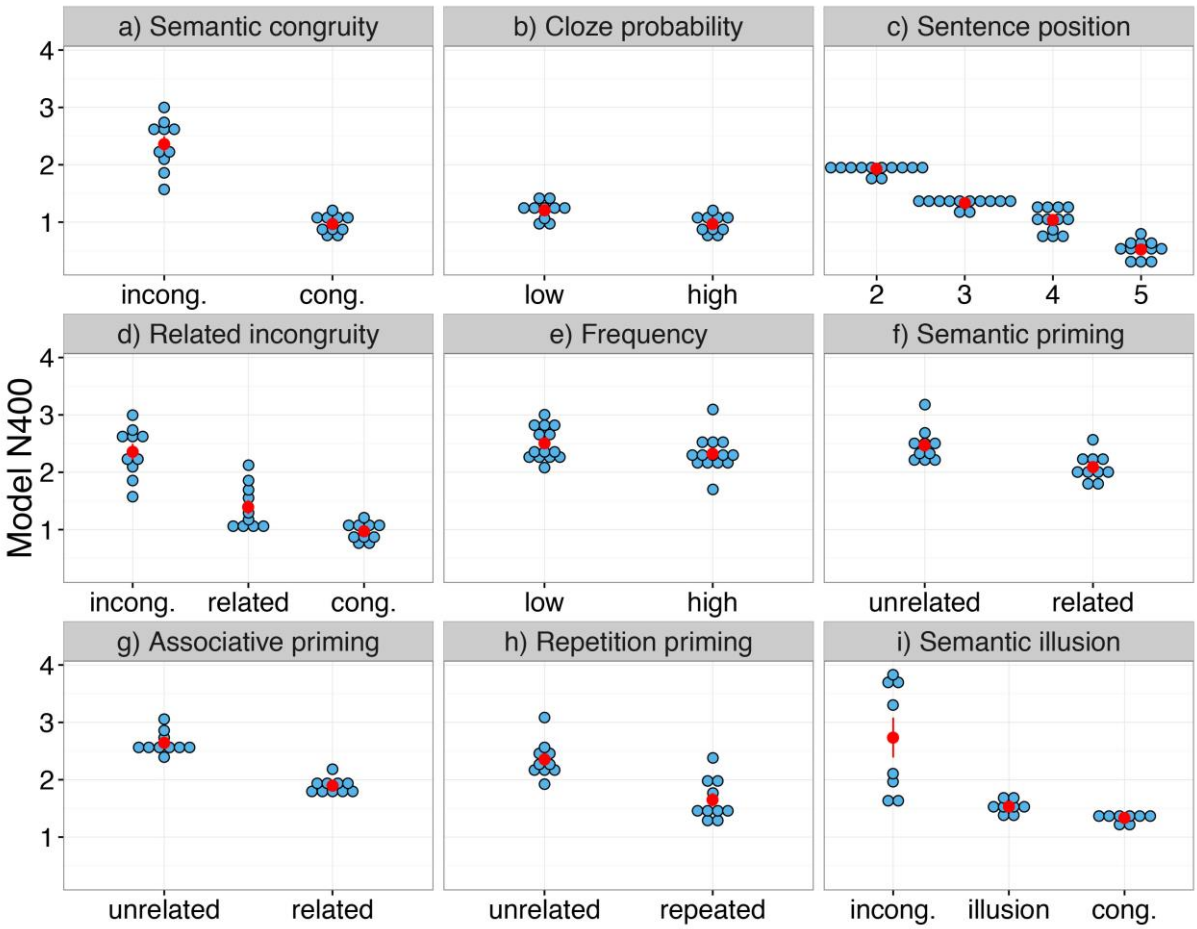


Figure 2. *Simulation results for the basic effects. Displayed is the model's N400 correlate, i.e. the update of the Sentence Gestalt layer activation – the model's probabilistic representation of sentence meaning - induced by the new incoming word. Cong., congruent; incong., incongruent. See text for details of each simulation. Each blue dot represents the results for one independent run of the model, averaged across items per condition; the red dots represent the means for each condition, and red error bars represent +/- SEM (sometimes invisible because bars may not exceed the area of the red dot). Statistical results (t_1 from the model analyses, t_2 from the item analyses): a, semantic incongruity: $t_{1(9)} = 25.00$, $p < .0001$, $t_{2(9)} = 11.24$, $p < .0001$; b, cloze probability: $t_{1(9)} = 8.56$, $p < .0001$, $t_{2(9)} = 6.42$, $p < .001$; c, position in sentence: $t_{1(9)} = 8.17$, $p < .0001$, $t_{2(11)} = 43.54$, $p < .0001$ from the second to the third sentence position; $t_{1(9)} = 4.73$, $p < .01$, $t_{2(11)} = 4.66$, $p < .01$, from the third to the fourth position; $t_{1(9)} = 17.15$, $p < .0001$, $t_{2(11)} = 12.65$, $p < .0001$, from the fourth to the fifth position; d, categorically related incongruities were larger than congruent, $t_{1(9)} = 10.63$, $p < .0001$, $t_{2(9)} = 3.31$, $p < .05$, and smaller than incongruent continuations, $t_{1(9)} = 14.69$, $p < .0001$, $t_{2(9)} = 12.44$, $p < .0001$; e, lexical frequency: $t_{1(9)} = 3.13$, $p < .05$, $t_{2(13)} = 3.26$, $p < .01$; f, semantic priming: $t_{1(9)} = 14.55$, $p < .0001$, $t_{2(9)} = 8.92$, $p < .0001$; g, associative priming: $t_{1(9)} = 14.75$, $p < .0001$, $t_{2(9)} = 18.42$, $p < .0001$; h, immediate repetition priming: $t_{1(9)} = 16.0$, $p < .0001$, $t_{2(9)} = 18.93$, $p < .0001$; i, semantic illusion: $t_{1(9)} = 2.09$, $p = .133$, $t_{2(7)} = 5.67$, $p < .01$, for the comparison between congruent condition and semantic illusion; $t_{1(9)} = 10.66$, $p < .0001$, $t_{2(7)} = 3.56$, $p < .05$, for the comparison between semantic illusion and incongruent condition.*



Supplementary Figure 1. *Simulation results for the basic effects (by item). Displayed is the model's N400 correlate, i.e. the update of the Sentence Gestalt layer activation – the model's probabilistic representation of sentence meaning - induced by the new incoming word. Cong., congruent; incong., incongruent. See text for details of each simulation. Here, each blue dot represents the results for one item, averaged across 10 independent runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM (sometimes invisible because bars may not exceed the area of the red dot). Statistical results are reported in the caption of Fig. 2 in the main text.*

Soon after the initial study it became clear that the N400 is graded, with larger amplitudes for acceptable sentence continuations with lower cloze probability (defined as the percentage of participants that continue a sentence fragment with that specific word in an offline sentence completion task), as in the example "Don't touch the wet *dog* (low cloze)*/ paint* (high cloze)"[10]. This result is also captured by the model: it exhibited larger SU for sentence endings presented with a low as compared to a high probability during training (Fig. 2b, Fig. 1, and Supplementary Fig. 1b). The graded character of the underlying process is further

supported empirically by the finding that N400s gradually decrease across the sequence of words in normal congruent sentences[11]. SU in the model correspondingly shows a gradual decrease across successive words in sentences (Fig. 2c and Supplementary Fig. 1c; see *online methods* for details).

Expectancy for words or semantic features? The findings discussed above would be consistent with the view that N400s reflect the inverse probability of a word in a specific context (i.e. word surprisal[12]), and indeed, a recent study observed a significant correlation between N400 and word surprisal measured at the output layer of a simple recurrent network (SRN) trained with a naturalistic corpus to predict the next word based on the preceding context[13]. However, there is evidence that N400s may not be a function of word probabilities *per se* but rather of probabilities of aspects of meaning signaled by words: N400s are smaller for incongruent completions that are closer semantically to the correct completion (and thus share more semantic features with it) than those that are semantically more distant. For example, consider the sentence: "They wanted to make the hotel look more like a tropical resort. So, along the driveway they planted rows of …*".* The N400 increase relative to *palms* (congruent completion) is smaller for *pines* (incongruent completion from the same basic level category as the congruent completion) than for *tulips* (incongruent completion not from the same basic level category as the congruent completion)"[14]. Our model captures these results: We compared SU for sentence completions that were presented with a high probability during training and two types of never-presented completions. SU was lowest for high probability completions, as expected; crucially, among never-presented completions, SU was smaller for those in the same semantic category as high probability completions compared to those that were categorically unrelated to all completions presented during training (Fig. 2d and Supplementary Fig. 1d).

Semantic integration versus lexical access? The sentence-level effects considered

above have often been taken to indicate that N400 amplitudes reflect the difficulty or effort required to integrate an incoming word into the preceding context[4,5]. However, a sentence context is not actually needed: N400 effects can also be obtained for words presented in pairs or even in isolation. Specifically, N400s are smaller for isolated words with a high as compared to a low lexical frequency[15]; for words (e.g. "bed") presented after a categorically related prime (e.g., "sofa") or an associatively related prime (e.g., "sleep") as compared to an unrelated prime[16]; and for an immediate repetition of a word compared to the same word following an unrelated prime[17]. Such N400 effects outside of a sentence context, especially the influences of repetition and lexical frequency, have led some researchers to suggest that N400 amplitudes do not reflect the formation of a representation of sentence meaning but rather lexical access to individual word meaning[3,18]. While the SG pattern probabilistically represents the meaning of a sentence if one is presented, the model can also process words presented singly or in pairs. Indeed, the model captures all four of the above-mentioned effects: First, SU was smaller for isolated words that occurred relatively frequently during training (Fig. 2e and Supplementary Fig. 1e). Furthermore, SU was smaller for words presented after words from the same semantic category as compared to words from a different category (Fig. 2f and Supplementary Fig. 1f), and smaller for words presented after associatively related words (objects presented after a typical action as in "chess" following "play") as compared to unrelated words (objects presented after an unrelated action as in "chess" following "eat") (Fig. 2g and Supplementary Fig. 1g). Finally, SU was smaller for immediately repeated words as compared to words presented after unrelated words (Fig. 2h and Supplementary Fig. 1h).

Semantic illusions and the N400. A finding that has puzzled the N400 community is the lack of a robust N400 effect in reversal anomalies (also termed *semantic illusions*): a surprisingly small N400 occurs in sentences such as "Every morning at breakfast, the eggs

would *eat*...". There is clearly an anomaly here – English syntactic conventions map eggs to the agent role despite the fact that eggs cannot eat – yet N400 amplitudes are only very slightly increased in such sentences as compared to the corresponding congruent sentences such as "Every morning at breakfast, the boys would *eat*..."[19]. This lack of a robust N400 effect in reversal anomalies is accompanied by an increase of the P600, a subsequent positive potential. In contrast, N400 but not P600 amplitudes are considerably larger in sentence variations such as "Every morning at breakfast, the boys would *plant*..."[19]. How can we understand this pattern? One analysis[20] treats these findings as challenging the view that the N400 is related to interpretation of sentence meaning, based on the argument that such sentences should produce a large N400 because they would require (for example) treating the eggs as the agents of eating, and this would require a substantial change in the meaning representation.

We find, however, that the semantic update in the SG model, which models the formation of a representation of sentence meaning, reproduces the pattern seen in the human N400 data. That is, the model exhibited only a very slight increase in SU for reversal anomalies (e.g., "At breakfast, the eggs *eat*…") as compared to typical continuations (e.g., "At breakfast, the man *eats…"*), and a substantial increase in SU for atypical continuations (e.g., "At breakfast, the man *plants*…") (Fig. 2i and Supplementary Fig. 1i). What happens in the SG model when it is presented with a reversal anomaly? Analysis of the query network's response to relevant probes (Fig. 3) suggests that the model exhibits a semantic illusion, in that the SG continues to implicitly represent the eggs as the patient instead of the agent of eating even after the word *eat* is presented. This observation is in line with the idea that, when presented with a reversal anomaly, comprehenders still settle at least initially into the most plausible semantic interpretation of the given input (i.e., the eggs being eaten) even if the sentence is anomalous syntactically[21].

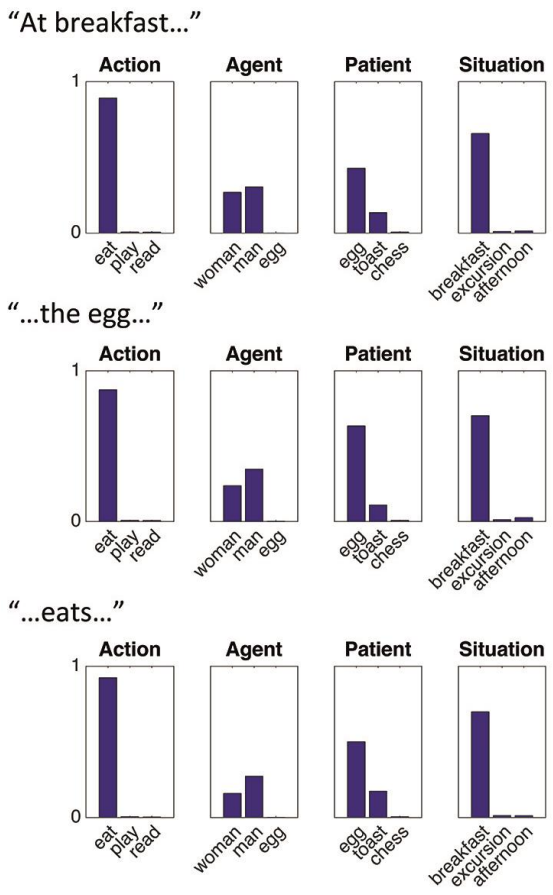


Figure 3. *Processing semantic illusions. Activation of selected output units while the model processes a sentence from the semantic illusion simulation: "At breakfast, the egg eats…". Note that the model continues to represent the egg as the patient (not the agent) of eating, even after the word "eat" has been presented, giving rise to a 'semantic illusion'.*

In summary, the model shows that the lack of an N400 increase for reversal anomalies is consistent with the view that the N400 reflects the updating of an implicit representation of sentence meaning. The model pre-dates the discovery of the semantic illusion phenomenon, and accounts for it without any modification, though the details of experience (for the model and for human learning) are expected to affect the size and nature of the update produced by particular anomalous sentences. Our account leaves open the possibility that other processes could potentially revise the initial interpretation. We consider how reversal anomalies might affect the P600 in the discussion.

Extensions

In all of the simulations above, it would have been possible to model the phenomena by treating the N400 as a direct reflection of change in estimates of event-feature probabilities, rather than as reflecting the update of an implicit internal representation of meaning that latently represents these estimates in a way that only becomes explicit when queried. In this section, we show that the implicit semantic update and the change in the networks' explicit estimates of feature probabilities in response to probes can pattern differently, with the implicit semantic update patterning more closely with the N400, supporting a role for the update of the learned implicit representation rather than explicit estimates of event-feature probabilities or objectively true probabilities in capturing neural responses (see *online methods* for details of these measures). We then consider how the implicit semantic update can drive connection-based learning in the update network, accounting for one final observed pattern of empirical findings.

Development. N400s change with increasing language experience and over developmental time. The examination of N400 effects in different age groups has shown that N400 effects increase with comprehension skills in babies[22] but later decrease with age[23,24]. A comparison of the effect of semantic congruity on SU at different points in training shows a developmental pattern consistent with these findings (Fig. 4, top, and Supplementary Fig. 2, top): the size of the congruity effect on SU first increased and then decreased as training proceeded. Interestingly, the decrease in the effect on SU over the second half of training was accompanied by a continuing increase in the effect of semantic congruity on the change in the model's explicit estimates of feature probabilities (Fig. 4, bottom, and Supplementary Fig. 2, bottom). This pattern indicates that, later in development, less change in activation at the SG layer is needed to effectively support larger changes in explicit probability estimates. This pattern is possible because the activation pattern at the SG layer does not explicitly represent

the probabilities of semantic features per se; instead it provides a basis (together with the connection weights in the query network) for estimating these probabilities when probed. As connection weights in the query network get stronger throughout the course of learning, smaller changes in SG activations are sufficient to produce big changes in output activations. This shift of labor from activation to connection weights is interesting in that it might underlie the common finding that neural activity often decreases as practice leads to increases in speed and accuracy of task performance[25].

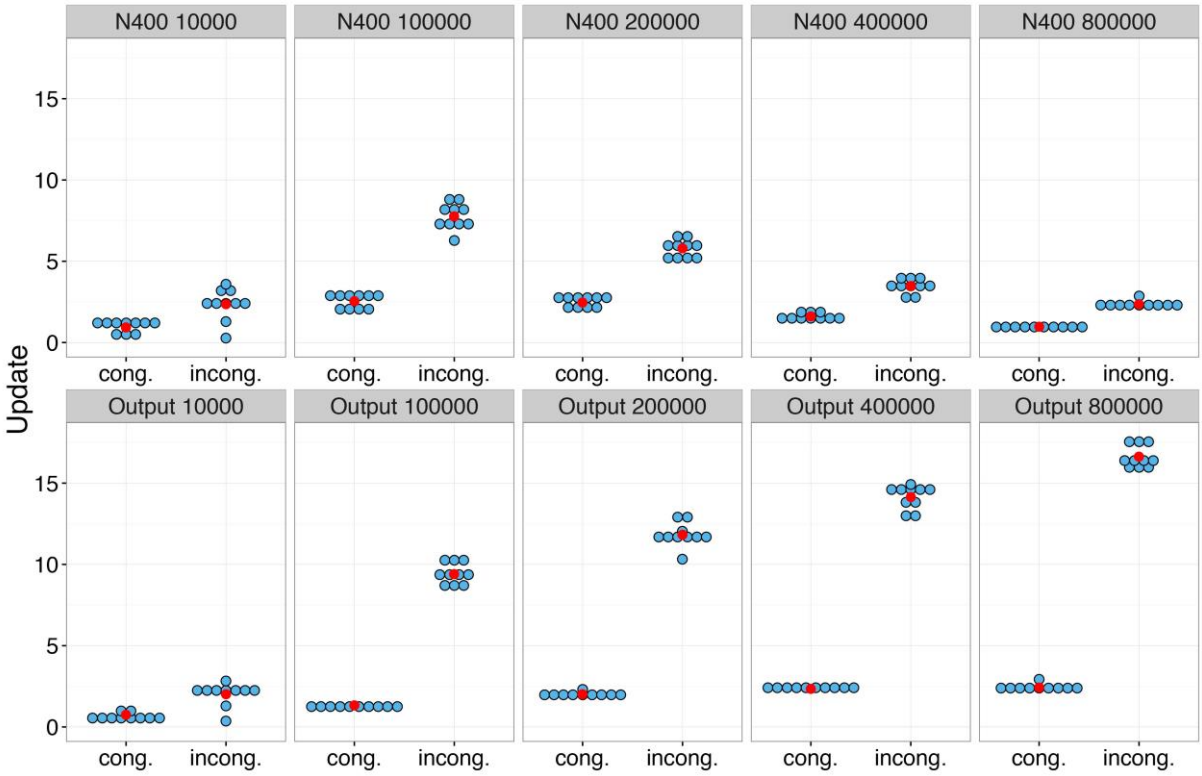
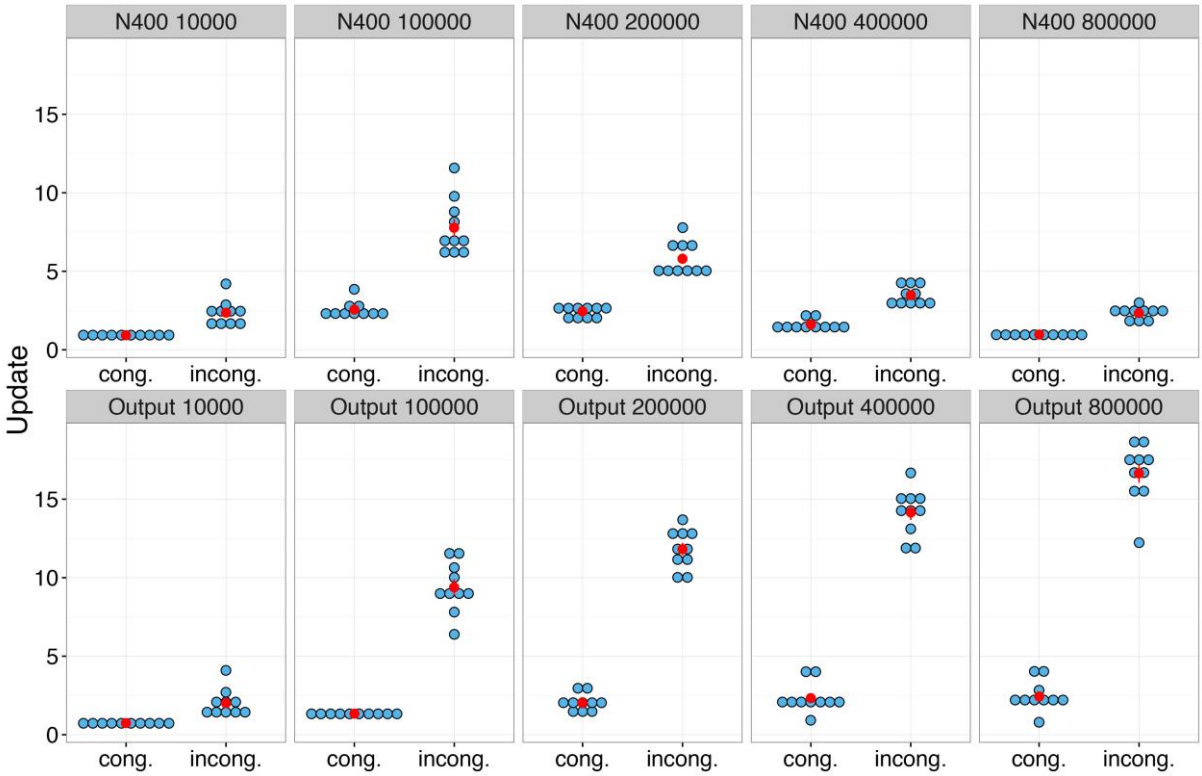


Figure 4. *Development across training. Semantic incongruity effects as a function of the number of sentences the model has been exposed to. Top. Semantic update at the model's hidden Sentence Gestalt layer shows at first an increase and later a decrease with additional training, in line with the developmental trajectory of the N400. Each blue dot represents the results for one independent run of the model, averaged across items per condition; the red dots represent the means for each condition, and red error bars represent +/- SEM. The size of the effect (i.e. the numerical difference between the congruent and incongruent condition) differed between all subsequent time points:* $t_{1(9)} = 17.02$, $p < .0001$, $t_{2(9)} = 6.94$, $p < .001$ *between 10000 and 100000 sentences;* $t_{1(9)} = 7.80$, $p < .001$, $t_{2(9)} = 10.05$, $p < .0001$ *between 100000 and 200000 sentences;* $t_{1(9)} = 14.69$, $p < .0001$, $t_{2(9)} = 6.87$, $p < .001$ *between 200000 and 400000 sentences;* $t_{1(9)} = 7.70$, $p < .001$, $t_{2(9)} = 3.70$, $p < .05$ *between 400000 and 800000 sentences. Bottom. Activation update at the output layer steadily increases with additional training, reflecting closer and closer approximation to the true conditional probability distributions embodied in the training corpus.*



Supplementary Figure 2. *Development across training (by item). Semantic incongruity effects as a function of the number of sentences the model has been exposed to. Top. Semantic update at the model's hidden Sentence Gestalt layer shows at first an increase and later a decrease with additional training, in line with the developmental trajectory of the N400. Each blue dot represents the results for one item, averaged across 10 independent runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM. Statistical results are reported in the caption of Fig. 4 in the main text. Bottom. Activation update at the output layer steadily increases with additional training, reflecting closer and closer approximation to the true conditional probability distributions embodied in the training corpus.*

Early sensitivity to a new language. A second language learning study showed robust influences of semantic priming on N400s while overt lexical decision performance in the newly trained language was still near chance[26]. We leave it to future work to do full justice to the complexity of second language learning, but as a first approximation we tested the model at a very early stage in training (Fig. 5a). Even at this early stage, SU was significantly influenced by semantic priming, associative priming, and semantic congruity in sentences (Fig. 5b and Supplementary Fig. 3) while overt estimates of feature probabilities were only weakly modulated by the words presented.

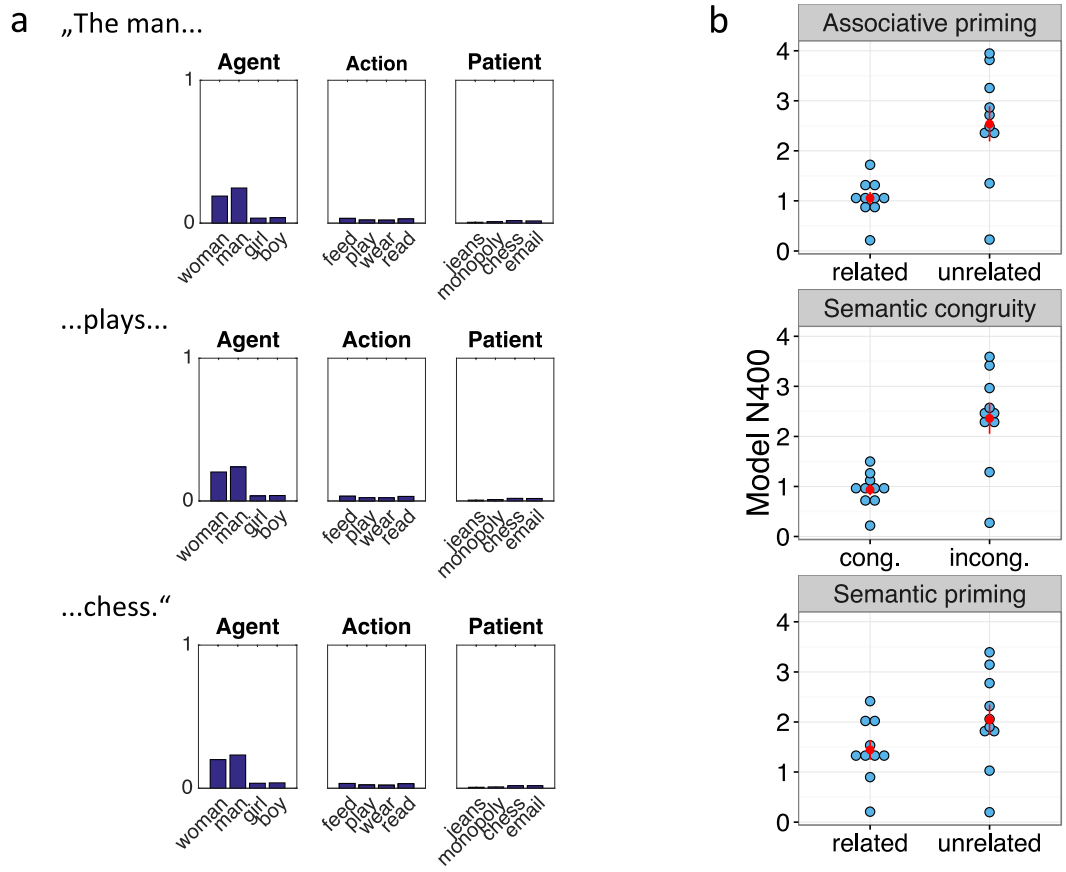
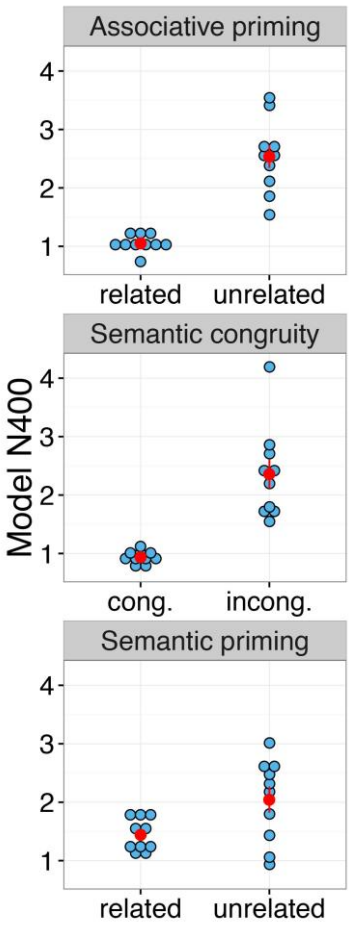


Figure 5. *Comprehension performance and semantic update effects at a very early stage in training. Cong., congruent; incong., incongruent. a. Activation of selected output units while the model is presented with the sentence "The man plays chess.". It can be seen that the model fails to activate the corresponding units at the output layer. The only thing that it has apparently learned at this point is which concepts correspond to possible agents, and it activates those in a way that is sensitive to their base rate frequencies (in the model's environment, woman and man are more frequent than girl and boy; see online methods), and with a beginning tendency to activate the correct agent ("man") most. b. Even at this low level of performance, there are robust effects of associative priming ($t_{1(9)} = 6.12$, $p < .001$, $t_{2(9)} = 7.31$, $p < .0001$, top), semantic congruity in sentences ($t_{1(9)} = 6.85$, $p < .0001$, $t_{2(9)} = 5.74$, $p < .001$, middle), and semantic priming ($t_{1(9)} = 5.39$, $p < .001$, $t_{2(9)} = 3.79$, $p < .01$, bottom), on the size of the semantic update, the model's N400 correlate. Each blue dot represents the results for one independent run of the model, averaged across items per condition; the red dots represent the means for each condition, and red error bars represent +/- SEM.*

Supplementary Figure 3 (see next page). *Comprehension performance and semantic update effects at a very early stage in training (by item). Cong., congruent; incong., incongruent. Even at a low level of performance (see Fig. 5a in the main text for illustration), there are robust effects of associative priming (top), semantic congruity in sentences (middle), and semantic priming (bottom). Here, each blue dot represents the results for one item, averaged across ten independent runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM. Statistical results are reported in the caption of Fig. 5 in the main text.*



The relationship between activation update and adaptation in a predictive system. The change induced by the next incoming word that we suggest underlies N400 amplitudes can be seen as reflecting the 'error' (difference or divergence) between the model's implicit probability estimates based on the previous word, and the updated estimate based on the next word in the sentence (see *online methods* for details). If the estimate after word *n* is viewed as a *prediction*, then this difference can be viewed as a kind of prediction error. It is often assumed that learning is based on such temporal difference or prediction errors[27–29] so that if N400 amplitudes reflect the update of a probabilistic representation of meaning, then larger N400s should be related to greater adaptation, i.e., larger adjustments to future estimates. Here we implement this idea, using the semantic update to drive learning. Importantly, this allows the model to learn just from listening or reading, when no separate event description is

provided. We then used this approach to simulate the finding that the effect of semantic incongruity on N400s is reduced by repetition: the first presentation of an incongruent completion, which induces larger semantic update compared to a congruent completion, leads to stronger adaptation, as reflected in a larger reduction in the N400 during a delayed repetition compared to the congruent continuation[30].

To simulate the observed interaction between repetition and semantic incongruity, we presented a set of congruent and incongruent sentences a first time, adapting the weights in the update network using the temporal difference signal on the SG layer to drive learning: The SG layer activation at the next word serves as the target for the SG layer activation at the current word, so that the error signal becomes $SG_{n+1} - SG_n$ (see *online methods*). We then presented all sentences a second time. Using this approach, we captured the greater reduction in the N400 with repetition of incongruent compared to congruent sentence completions (Fig. 6 and Supplementary Fig. 4).

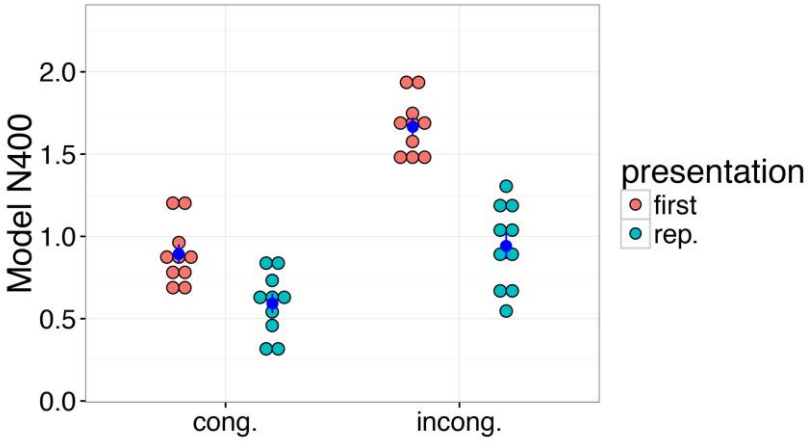
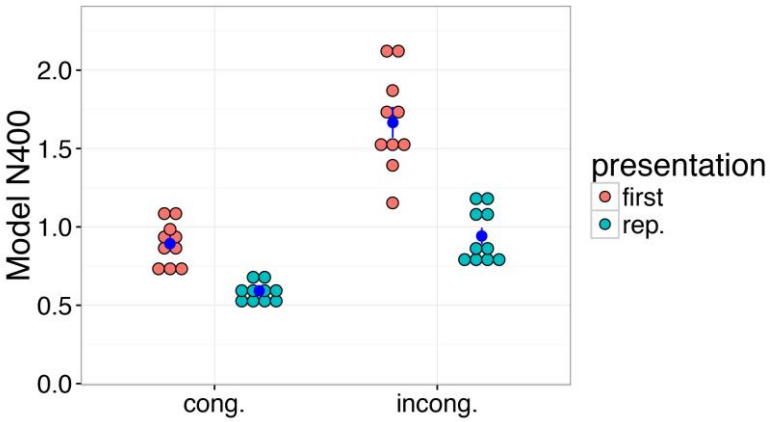


Figure 6. *Simulation of the interaction between delayed repetition and semantic incongruity. Cong., congruent; incong., incongruent; rep., repeated. Each red or green dot represents the results for one independent run of the model, averaged across items per condition; the blue dots represent the means for each condition, and blue error bars represent +/- SEM. There were significant main effects of congruity, $F_1(1,9) = 214.13$, $p < .0001$, $F_2(1,9) = 115.66$, $p < .0001$, and repetition, $F_1(1,9) = 48.47$, $p < .0001$, $F_2(1,9) = 109.78$, $p < .0001$, and a significant interaction between both factors, $F_1(1,9) = 83.30$, $p < .0001$, $F_2(1,9) = 120.86$, $p < .0001$; post-hoc comparisons showed that even though the repetition effect was larger for incongruent as compared to congruent sentence completions, it was significant in both conditions, $t_{1(9)} = 4.21$, $p < .01$, $t_{2(9)} = 6.90$, $p < .0001$, for the congruent completions, and $t_{1(9)} = 8.78$, $p < .0001$, $t_{2(9)} = 12.02$, $p < .0001$, for the incongruent completions.*



Supplementary Figure 4. *Simulation of the interaction between delayed repetition and semantic incongruity (by item). Each red or green dot represents the results for one item, averaged across 10 runs of the model; blue dots represent means for each condition, and blue error bars represent +/- SEM. Statistical results are reported in the caption of Fig. 6.*

Notably, the summed magnitude of the signal that drives learning corresponds exactly to our N400 correlate, highlighting the relationship between semantic update, prediction error, and experience-driven learning. Thus, our account predicts that in general, larger N400s should induce stronger adaptation. Though further investigation is needed, there is some evidence consistent with this prediction: larger N400s to single word presentations during a study phase have been shown to predict enhanced implicit memory (measured by stem completion in the absence of explicit memory) during test[31].

Discussion

The N400 ERP component is widely used to investigate the neurocognitive processes underlying the processing of meaning in language, but the brain basis of meaning construction is not yet fully understood. Here, we advance the view that N400 amplitudes reflect the change induced by an incoming stimulus in an implicit probabilistic representation of meaning. We offer an explicit formulation of the role of such a representation in language processing in which connection-based knowledge that supports this representation is shaped by experience with language occurring both in the context of observed events and when language is processed in isolation.

The pattern of activation in the model's Sentence Gestalt (SG) layer latently predicts the semantic attributes of the entire event described by a sentence, capturing base-rate probabilities (before sentence processing begins) and adjusting this pattern of activation as each word of the sentence is presented. It is important to note that this kind of prediction does not refer to explicit intentional prediction of specific items but rather to a general configuration of the system in the sense that the model (and presumably the brain) becomes tuned through experience to anticipate likely upcoming input to respond to it with little additional effort. This entails that semantic activation changes induced by new incoming input as revealed in the N400 reflect the discrepancy between probabilistically anticipated and encountered information about aspects of sentence meaning and at the same time correspond to the learning signal driving adaptation of connection-based knowledge representations. In this sense, our approach conceptually overlaps with predictive coding[28]. Our simulations suggest that the semantic system may not represent aspects of meaning explicitly but rather uses a summary representation that supports explicit probability estimates when queried, becoming more and more efficient as learning progresses.

Recently, other studies have also begun to link the N400 to computational models. Most of these have concentrated on words presented singly or after a preceding prime, and therefore do not address processing in a sentence context[18,32,33]. Two modeling studies focus on sentence processing. One of these studies observed a correlation between N400s and word surprisal as estimated by a simple recurrent network (SRN) trained to predict the next word based on the preceding context[13]. Because the SRN's predictions generalize across contexts and are mediated by a similarity-based internal representation, it can potentially account for effects of semantic similarity on word surprisal, and would thus share some predictions with the SG model. However, an account of N400s in terms of word surprisal faces some difficulties. First, word surprisal reflects both semantic and syntactic expectation violations, while the N400 is specific to meaning processing and syntactic violations typically modulate

different ERPs[34]. Indeed, the observed correlation between surprisal in the SRN and N400 held only for content words and not for function words[13]. Furthermore, the SRN may have difficulty accounting for the decrease of N400 effects with age (because surprisal is measured in terms of the estimates of word probabilities, which become sharper as learning progresses), as well as the small N400 in reversal anomalies: When presented with "At breakfast, the eggs would *eat*…", word surprisal would likely be large, while semantic update in the SG model shows only a very slight increase, in line with N400 data[19].

The other sentence-level model focuses specifically on reversal anomalies, assuming separate lexical and semantic integration stages[20], based on the assumption that reversal anomalies must produce a large update in a representation of sentence meaning. In this model, change in lexical activation (which is small in reversal anomalies due to priming, e.g. from *breakfast* and *eggs* to *eat*) is linked to the N400; the change in activation representing sentence meaning is assigned to the later, P600 ERP component. However, as discussed above (see *Results*) our model accounted for the small size of the N400 in reversal anomalies without separate mechanisms for lexical access and semantic interpretation, and addresses a wider range of N400 effects. Furthermore, variables which should influence the amount of change in a representation of sentence meaning, such as cloze probability or surprise, consistently influence N400 but not necessarily P600 amplitudes[10,13], supporting our view that the N400, not the P600, corresponds to the update of a representation of sentence meaning. The basis of the P600 requires further investigation; it might reflect detection of the anomaly, consistent with the proposal that the P600 is an instance of the oddball-sensitive P3[35]. In that case, detection of the anomaly might trigger processes that could ultimately revise the semantic interpretation of the sentence.

The current work opens up an opportunity for extensive further investigations, addressing a wide range of behavioral as well as neural aspects of sentence processing. The model's query language and training corpus will need to be extended to address the full range

of relevant phenomena, and it will need to be integrated into a more complete account of the neuro-mechanistic processes that take place during language processing, including the P600 and other ERP components as well as signals that have been detected using other measurement modalities[34,36]. While extending the model will be worthwhile, it nevertheless makes a useful contribution to understanding the brain processes underlying language comprehension in its current simple form. A recent comprehensive review on the N400 ERP component[2] concluded that the N400 "does not readily map onto specific subprocesses posited in traditional frameworks" (p. 18) and that therefore none of the available accounts of N400 amplitudes - proposing functional localizations at some specific point along a processing stream from prelexical analysis over lexical processing to word recognition, semantic access, and semantic integration - could explain the full range of N400 data. Instead, the authors suggest that N400 amplitudes might best be understood as a "temporally delimited electrical snapshot of the intersection of a feedforward flow of stimulus-driven activity with a state of the distributed, dynamically active neural landscape that is semantic memory." (p. 21). This view seems reminiscent of the SG model, in which incoming stimuli serve as 'cues to meaning'[37] which change the overall semantic representation of the described event. Crucially, the model provides a computationally explicit account of the nature and role of this distributed representation and how it changes through stimulus-driven activity as meaning is dynamically constructed during comprehension. This simple model's successes in capturing a diverse body of empirically observed neural responses suggest that the principles of semantic representation and processing it embodies may capture essential aspects of human language comprehension.

Online Methods

We begin by describing the implicit probabilistic theory of meaning underlying the Sentence Gestalt model and relate the updates in the model to other probabilistic measures of surprise. Next we describe the new semantic update driven learning rule used in simulating the reduction in the incongruity effect due to repetition. We then provide details on the model's training environment as well as the protocols used for training the model and for the simulations of empirical findings. Figure 1 in the main text presents the network architecture and the processing flow in the model.

Implicit probabilistic theory of sentence meaning

The theory of meaning embodied in the Sentence Gestalt model holds that sentences constrain an implicit probabilistic representation of the meanings speakers intend to convey through these sentences. The representation is implicit in that no specific form for the representation is prescribed, nor are specific bounds set on the content of the representation of meaning. Instead, sentences are viewed as conveying information about situations or events, and a representation of meaning is treated as a representation that provides the comprehender with a basis for estimating the probabilities of aspects of the situation or event the sentence describes. To capture this we characterize the ensemble of aspects as an ensemble of queries about the event, with each query associated with an ensemble of possible responses. In the general form of the theory, the queries could range widely in nature and scope (encompassing, for example, whatever the comprehender should expect to observe via any sense modality or subsequent linguistic input, given the input received so far). In implementations to date, at least four different query formats have been considered[38–40], including a natural language-based question and answer format (Fincham & McClelland, 1997, Abstract). Queries may also vary in their probability of being posed (hereafter called *demand probability*), and the correct answer to a particular query may be uncertain, since sentences may be ambiguous,

vague or incomplete. A key tenet of the theory is that aspects of meaning can often be estimated without being explicitly described in a sentence, due to knowledge acquired through past experience[38]. If events involving cutting steak usually involve a knife, the knife would be understood, even without ever having been explicitly mentioned in a sentence.

The theory envisions that sentences are uttered in situations where information about the expected responses to a probabilistic sample of queries is often available to constrain learning about the meaning of the sentence. When such information is available, the learner is thought to be (implicitly) engaged in attempting to use the representation derived from listening to the sentence to anticipate the expected responses to these queries and to use the actual responses provided with the queries to bring the estimates of the probabilities of these responses in line with their probabilities in the environment. This process is thought to occur in real time as the sentence unfolds; for simplicity it is modeled as occurring word by word as the sentence is heard.

As an example, consider the sequence of words 'The man eats' and the query, 'What does he eat'? What the theory assumes is that the environment specifies a probability distribution over the possible answers to this and many other questions, and the goal of learning is to form a representation that allows the comprehender to match this probability distribution.

More formally, the learning environment is treated as producing sentence-event-description pairs according to a probabilistic generative model. The sentence consists of a sequence of words, while the event-description consists of a set of queries and associated responses. Each such pair is called an *example.* The words in the sentence are presented to the neural network in sequence, and after each word, the system can be probed for its response to each query, which is conditional on the words presented so far (we use w_n to denote the sequence of words up to and including word n). The goal of learning is to minimize the expected value over the distribution of examples of a probabilistic measure (the Kullback-

Leibler divergence, D_{KL}) of the difference between the distribution of probabilities p over possible responses r to each possible query and the model's estimates ρ of the distribution of these probabilities, summed over all of the queries q occurring after each word, and over all of the words in the sentence. In this sum, the contribution of each query is weighed by its demand probability conditional on the words seen so far, represented $p(q|w_n)$. We call this the *expected value E of the summed divergence measure*, written as:

$$E\left(\sum_n \sum_q p(q|w_n)\, D_{KL}(p(r|q,w_n)||\rho(r|q,w_n))\right)$$

In this expression the divergence for each query, $D_{KL}(p(r|q,w_n)||\rho(r|q,w_n))$, is given by

$$\sum_r p(r|q,w_n) \log\left(\frac{p(r|q,w_n)}{\rho(r|q,w_n)}\right)$$

It is useful to view each combination of a query q and sequence of words w_n as a context, henceforth called C. The sequence of words 'the man eats' and the query 'what does he eat?' is an example of one such context. To simplify our notation, we will consider each combination of q and w_n as a context C, so that the divergence in context C, written $D_{KL}(C)$, is $\sum_r p(r|C) \log\left(\frac{p(r|C)}{\rho(r|C)}\right)$. Note that $D_{KL}(C)$ equals 0 when the estimates match the probabilities (that is, when $p(r|C) = \rho(r|C)$ for all r) in context C, since log(x/x) = log(1) = 0. Furthermore, the expected value of the summed divergence measure is 0 if the estimates match the probabilities for all C.

Because the real learning environment is rich and probabilistic, the number of possible sentences that may occur in the environment is indefinite, and it would not in general be possible to represent the estimates of the conditional probabilities explicitly (e.g. by listing them in a table). A neural network solves this problem by providing a mechanism that can process any sequence of words and associated queries that are within the scope of its environment, allowing it to generate appropriate estimates in response to queries about sentences it has never seen before[38].

Learning occurs from observed examples by stochastic gradient descent: A training example consisting of a sentence and a corresponding set of query-response pairs is drawn from the environment. Then, after each word of the sentence is presented, each of the queries is presented along with the response that is paired with it in the example. This response is treated as the target for learning, and the model adjusts its weights to increase its probability of giving this response under these circumstances. This procedure tends to minimize the expected value of the summed divergence measure over the environment, though the model's estimates will vary around the true values in practice as long as a non-zero learning rate is used. In that case the network will be sensitive to recent history and can gradually change its estimates if there is a shift in the probabilities of events in the environment.

The implemented query-answer format and standard network learning rule

In the implementation of the model used here, the queries presented with a given training example can be seen as questions about attributes of the possible fillers of each of a set of possible roles in the event described by the sentence. There is a probe for each role, which can be seen as specifying a set of queries, one for each of the possible attributes of the filler of the role in the event. For example, the probe for the agent role can be thought of as asking, in parallel, a set of binary yes-no questions, one about each of several attributes or features *f* of the agent of the sentence, with the possible responses to the question being 1 (for yes the feature is present) or 0 (the feature is not present). For example, one of the features specifies whether or not the role filler is male. Letting $p(v|f,C)$ represent the probability that the feature has the value *v* in context *C* (where now context corresponds to the role being probed in the training example after the *n*th word in the sentence has been presented), the divergence can be written as $\sum_{v=1,0} p(v|f,C) \log\left(\frac{p(v|f,C)}{\rho(v|f,C)}\right)$. Writing the terms of the sum explicitly, this becomes $p(1|f,C) \log\left(\frac{p(1|f,C)}{\rho(1|f,C)}\right) + p(0|f,C) \log\left(\frac{p(0|f,C)}{\rho(0|f,C)}\right)$. Using the fact that the two possible answers are mutually exclusive and exhaustive, the two probabilities

must sum to 1, so that $p(0|f,C) = 1 - p(1|f,C)$; and similarly, $\rho(0|f,C) = 1 - \rho(1|f,C)$. Writing $p(f|C)$ as shorthand for $p(1|f,C)$ and $\rho(f|C)$ for $\rho(1|f,C)$, and using the fact that $\log(a/b) = \log(a) - \log(b)$ for all a,b, the expression for $D_{KL}(f,C)$ becomes

$$\big(p(f|C)\log\big(p(f|C)\big) + \big(1 - p(f|C)\big)\log\big(1 - p(f|C)\big)\big)$$

$$-\big(p(f|C)\log\big(\rho(f|C)\big) + (1 - p(f|C))\log\big(1 - \rho(f|C)\big)\big)$$

The first part of this expression contains only environmental probabilities and is constant, so that minimizing the expression as a whole is equivalent to minimizing the second part, called the *cross-entropy* $CE(f,C)$ between the true and the estimated probability that the value of feature $f = 1$ in context C:

$$CE(f,C) = -\big(p(f|C)\log\big(\rho(f|C)\big) + (1 - p(f|C))\log\big(1 - \rho(f|C)\big)\big)$$

The goal of learning is then to minimize the sum of this quantity across all features and situations.

The actual value of the feature for a particular role in a randomly sampled training example e is either 1 (the filler of the role has the feature) or 0 (the filler does not have the feature). This actual value is the target value used in training, and is represented as $t(f|C_e)$, where we use C_e to denote the specific instance of this context in the training example (note that the value of a feature depends on the probed role in the training example, but stays constant throughout the processing of each of the words in the example sentence). The activation a of a unit in the query network in context C_e, $a(f|C_e)$, corresponds to the network's estimate of the probability that the value of this feature is 1 in the given context; we use a instead of ρ to call attention to the fact that the probability estimates are represented by unit activations. The *cross-entropy* between the target value for the feature and the probability estimate produced by the network in response to the given query after word n then becomes:

$$CE(f,C_e) = -(t(f|C_e)\log(a(f|C_e)) + (1 - t(f|C_e))\log(1 - a(f|C_e)))$$

To see why this expression represents a sample that can be used to estimate *CE*(*f*,*C*) above, it is useful to recall that the value of a feature in a given context varies probabilistically across training examples presenting this same context. For example, for the context 'the man eats …', the value of a feature of the filler of the patient role can vary from case to case. Over the ensemble of training examples, the probability that $t(f|C_e) = 1$ corresponds to $p(f|C)$, so that the expected value of $t(f|C_e)$ over a set of such training examples will be $p(f|C)$, and the average value of $CE(f,C_e)$ over such instances will approximate $CE(f,C)$.

Now, the network uses units whose activation *a* is given by the logistic function of its net input, such that $a = 1/(1 + e^{-net})$, where the net input is the sum of the weighted influences of other units projecting to the unit in question, plus its bias term. As has long been known[41], the negative of the gradient of this cross-entropy measure with respect to the net input to the unit is simply $t(f|C_e) - a(f|C_e)$. This is the signal back-propagated through the network for each feature in each context during standard network training (see section *simulation details/ training protocol* for more detail).

Probabilistic measures of the surprise produced by the occurrence of a word in a sentence

Others have proposed probabilistic measures of the surprise produced by perceptual or linguistic inputs[9,12]. In the framework of our approach to the characterization of sentence meaning, we adapt one of these proposals[9], and use it to propose measures of three slightly different conceptions of surprise: The normative surprise, the subjective explicit surprise, and the implicit surprise – the last of which corresponds closely to the measure we use to model the N400.

We define the normative surprise (NS) resulting from the occurrence of the *n*th word in a sentence *s* as the KL divergence between the environmentally determined distribution of responses *r* to the set of demand-weighted queries *q* before and after the occurrence of word w_n:

$$NS(w_n) = \sum_q p(q|w_n) \sum_{r|q,s} p(r|q,w_n)\ log\left(\frac{p(r|q,w_n)}{p(r|q,w_{n-1})}\right)$$

If one knew the true probabilities, one could calculate the normative surprise and attribute it to an ideal observer. In the case where the queries are binary questions about features as in the implemented version of the SG model this expression becomes:

$$NS(w_n) = \sum_q p(q|w_n)\ \Big(p(f|q,w_n)\log\Big(\frac{p(f|q,w_n)}{p(f|q,w_{n-1})}\Big) + \big(1-p(f|q,w_n)\big)\log\Big(\frac{1-p(f|q,w_n)}{1-p(f|q,w_{n-1})}\Big)\Big)$$

To keep this expression simple, we treat q as ranging over the features of the fillers of all of the probed roles in the sentence.

The explicit subjective surprise ESS treats a human participant or model thereof as relying on subjective estimates of the distribution of responses to the set of demand-weighted queries. In the model these are provided by the activations *a* of the output units corresponding to each feature:

$$ESS(w_n) = \sum_q \rho(q|w_n)\ \Big(a(f|q,w_n)\log\Big(\frac{a(f|q,w_n)}{a(f|q,w_{n-1})}\Big) + \big(1-a(f|q,w_n)\big)\log\Big(\frac{1-a(f|q,w_n)}{1-a(f|q,w_{n-1})}\Big)\Big)$$

Our third measure, the implicit surprise (IS) is a probabilistically interpretable measure of the change in the pattern of activation over the learned internal meaning representation (corresponding to the SG layer in the model). Since the unit activations are

constrained to lie in the interval between 0 and 1, they can be viewed intuitively as representing estimates of probabilities of implicit underlying meaning dimensions or *microfeatures*[42] that together constrain the model's estimates of the explicit feature probabilities. In this case we can define the implicit surprise as the summed KL divergence between these implicit feature probabilities before and after the occurrence of word *n*, using a_i to represent the estimate of the probability that the feature characterizes the meaning of the sentence and $(1 - a_i)$ to represent the negation of this probability:

$$IS(w_n) = \sum_i \left(a_i(w_n) \log \left(\frac{a_i(w_n)}{a_i(w_{n-1})} \right) + \left(1 - a_i(w_n)\right) \log \left(\frac{1 - a_i(w_n)}{1 - a_i(w_{n-1})} \right) \right)$$

The actual measure we use for the semantic update (SU) as defined in the main text is similar to the above measure in being a measure of the difference or divergence between the activation at word *n* and word *n-1*, summed over the units in the SG layer:

$$SU(w_n) = \sum_i |a_i(w_n) - a_i(w_{n-1})|$$

The SU and IS are highly correlated and have the same minimum (both measures are equal to 0 when the activations before and after word *n* are identical). We use the analogous measure over the outputs of the query network, called the explicit subjective update (ESU) to compare to the SU in the developmental simulation reported in the main text:

$$ESU(w_n) = \sum_q \rho(q|w_n) \, |a(f|q, w_n) - a(f|q, w_{n-1})|$$

As before we treat q as ranging over all of the features of the fillers of all of the probed roles in the sentence. In calculating the ESU or the ESS, the queries associated with the presented sentences are all used, with $\rho(q|w_n) = 1$ for each one.

The simulation results presented in the main text show the same pattern in all cases if the ESS and IS are used rather than the SU and ESU.

Semantic update driven learning rule

The semantic update driven learning rule introduced in this article for the Sentence Gestalt model is motivated by the idea that later-coming words in a sentence provide information that can be used to teach the network to optimize the probabilistic representation of sentence meaning it derives from words coming earlier in the sentence. We briefly consider how this idea could be applied to generate signals for driving learning in the query network, in a situation where the teaching signal (in the form of a set of queries and corresponding feature values) corresponding to the actual features of an event are available to the model only after the presentation of the last word of the sentence (designated word *N*). In that situation, the goal of learning for the last word can be treated as the goal of minimizing the KL divergence between the outputs of the query network after word *N* and the target values of the features of the event $t(f|q,e)$. As in the standard learning rule, this reduces to the cross-entropy, which for a single feature is given by

$$CE(f, q, w_N) = -\big(t(f|q,e)\log\big(a(f|q,w_N)\big) + (1 - t(f|q,e))\log\big(1 - a(f|q,w_N)\big)\big)$$

A single {*sentence*, *event*} pair chosen from the environment would then provide a sample from this distribution. As is the case in the standard training regime, the negative of the gradient with respect to the net input to a given output feature unit in the query network after a given probe is simply $t(f|q,e) - a(f|q,w_N)$. This is then the error signal propagated back through the network. To train the network to make better estimates of the feature

probabilities from the next to last word in the sentence (word *N*-1), we can use the difference between the activations of the output units after word *N* as the teaching signal for word *N*-1, so for a given feature unit the estimate of the gradient with respect to its net input simply becomes $a(f|q,w_N) - a(f|q,w_{N-1})$. Using this approach, as $a(f|q,w_N)$ comes to approximate $t(f|q,e)$ it thereby comes to approximate the correct target for $a(f|q, N\text{-}1)$. This cycle repeats for earlier words, so that as $a(f|q, N\text{-}1)$ comes to approximate $a(f|q, N)$ and therefore $t(f|q, e)$ it also comes to approximate the correct teacher for $a(f|q, N\text{-}2)$, etc. This approach is similar to the temporal difference (TD) learning method used in reinforcement learning[43] in situations where reward becomes available only at the end of an episode, except that here we would be learning the estimates of the probabilities for all of the queries rather than a single estimate of the final reward at the end of an episode. This method is known to be slow and can be unstable, but it could be used in combination with learning based on episodes in which teaching information is available throughout the processing of the sentence, as in the standard learning rule for the SG model.

The semantic update based learning rule we propose extends the idea described above, based on the observation that the pattern of activation over the SG layer of the update network serves as the input pattern that allows the query network to produce estimates of probabilities of alternative possible responses to queries after it has seen some or all of the words in a sentence. Consider for the moment an ideally trained network in which the presentation of each word produces the optimal update to the SG representation based on the environment it had been trained on so far, so that the activations at the output of the query network would correspond exactly to the correct probability estimates. Then using the SG representation after word n+1 as the target for training the SG representation after word n would allow the network to update its implicit representation based on word n to capture changes in the environmental probabilities as these might be conveyed in a sentence. More formally, we propose that changing the weights in the update network to minimize the Implicit Surprise

allows the network to make an approximate update to its implicit probabilistic model of sentence meaning, providing a way for the network to learn from linguistic input alone. The negative of the gradient of the Implicit Surprise with respect to the net input to SG unit *i* after word *n* is given by $a_i(w_n) - a_i(w_{n-1})$. This is therefore the signal that we back propagate through the update network to train the connections during implicit temporal difference learning. As noted in the main text, the sum over the SG units of the absolute value of this quantity also corresponds to the SU, our model's N400 correlate.

Simulation Details

Environment. The model environment consists of {sentence, event} pairs probabilistically generated online during training according to constraints embodied in a simple generative model (see Fig. 7a). The sentences are single clause sentences such as "At breakfast, the man eats eggs in the kitchen". They are stripped of articles as well as inflectional markers of tense, aspect, and number, and are presented as a sequence of constituents, each consisting of a content word and possibly one closed class word such as a preposition or passive marker. A single input unit is dedicated to each word in the model's vocabulary. In the example above, the constituents are "at breakfast", "man", "eats", "eggs", "in kitchen", and presentation of the first constituent corresponds to activating the input units for "at" and "breakfast".

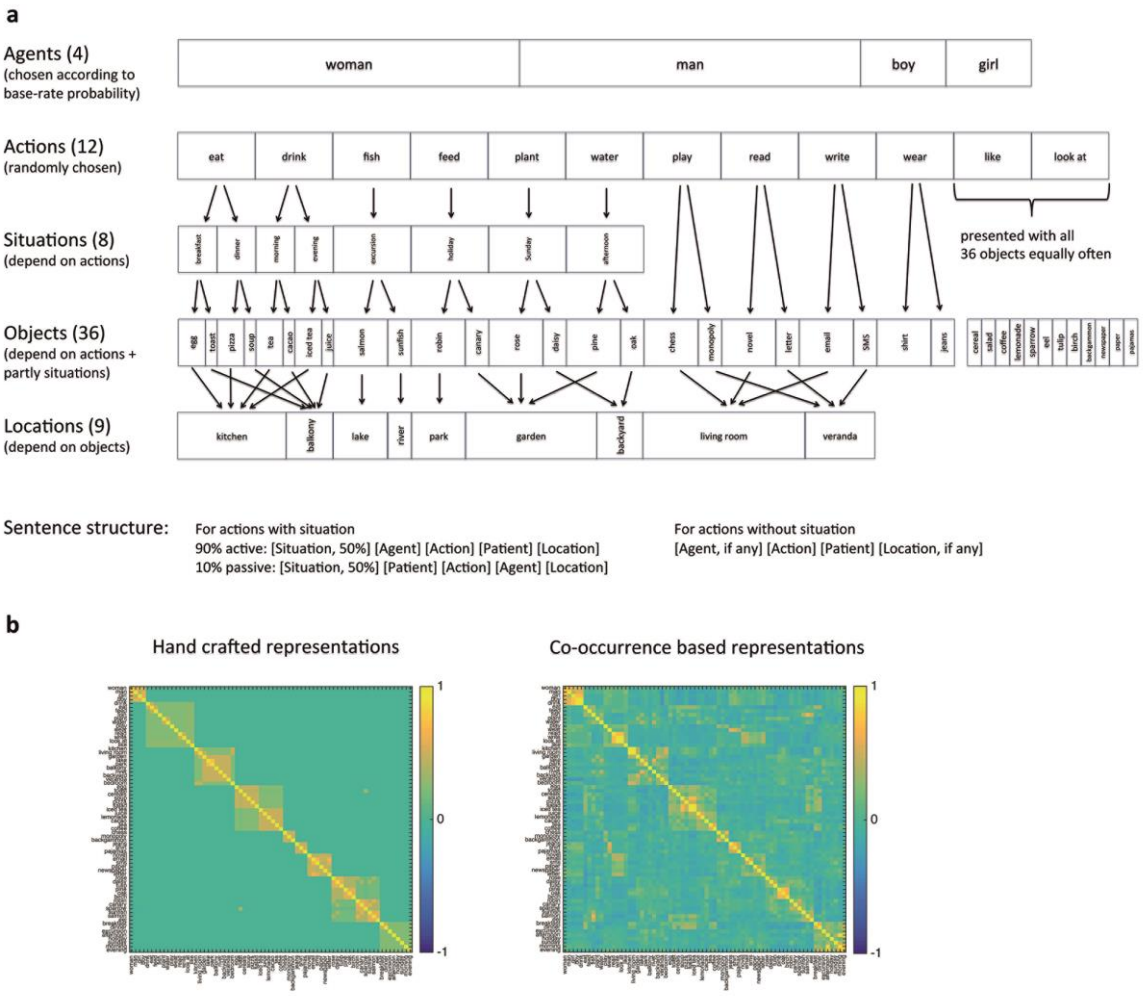


Figure 7. a. The sentence/ event generator used to train the model. Bar width corresponds to relative probability. First, one out of twelve actions is chosen with equal probability. Then, for every action except one ("look at") an agent is chosen ("woman" and "man" each with a probability of .4, "boy" and "girl" with a probability of .1). Next, a situation is chosen depending on the action. Some actions can occur in two possible situations, some in one, and some without a specified situation. Even if an action occurs in a specific situation, the corresponding word is presented only with a probability of .5 in the sentence while the situation is always part of the event representation. Then, depending on the action (and in the case that an action can occur in two possible situations, depending on the situation) an object/patient is chosen. For each action or situation (except for "like" and "look at" for which all 36 objects are chosen equally often) there is a high probability and a low probability object (if the agent is "man" or "woman", the respective high/low probabilities are .7/.3, if the agent is "girl" or "boy", the probabilities are .6/.4). The high and low probability objects occurring in the same specific action context are always from the same semantic category, and for each category, there is a third object which is never presented in that action context and instead only occurs in the unspecific "like" or "look at" contexts (to enable the simulation of categorically related incongruities; these are the twelve rightmost objects in the figure; here bar width is larger than probability to maintain readability). Possible sentence structures are displayed below. b. Similarity matrices of the hand-crafted semantic representations used for the current model (left) and representations based on a principal component analysis on word vectors derived from co-occurrences in large text corpora[44]. The correlation between the matrices is r = .73.

The events are characterized as sets of role filler pairs, in this case: agent – man, action – eat, patient – eggs, location – kitchen, situation - breakfast. Each thematic role is represented by a single unit at the probe and output layer. For the filler concepts, we used feature-based semantic representations such that each concept was represented by a number of units (at the probe and output layer) each corresponding to a semantic feature. For instance, the concept "daisy" was represented by five units. The units have labels that allow the reader to keep track of their roles but the model is not affected by the labels themselves, only by the similarity relationships induced by these labels. For example, the semantic features of "daisy" are labeled "can grow", "has roots", "has petals", "yellow", and "daisy". The feature-based representations were handcrafted to create graded similarities between concepts roughly corresponding to real world similarities as in other models of semantic representation[45,46]. For instance, all living things shared a semantic feature ("can grow"), all plants shared an additional feature ("has roots"), all flowers shared one more feature ("has petals") and then the daisy had two individuating features ("yellow" and its name "daisy") so that the daisy and the rose shared three of their five semantic features, the daisy and the pine shared two features, the daisy and the salmon shared only one feature, and the daisy and the email did not share any features (see the Supplementary Table 1 for a complete list of concepts and features). Comparison of a similarity matrix of the concepts based on our hand-crafted semantic representations and representations based on a principal component analysis (PCA) performed on semantic word vectors derived from co-occurrences in large text corpora[44] showed a reasonable correspondence (r = .73; see Fig. 7b), suggesting that the similarities among the hand-crafted conceptual representations roughly matched real world similarities (as far as they can be derived from co-occurrence statistics).

Training protocol. The training procedure is intended to approximate a situation in which a language learner has observed an event and thus has a complete representation of the event available, and then hears a sentence about it so that learning can be based on a

comparison of the current output of the comprehension mechanism and the event. Furthermore, it implements the assumption that listeners anticipate the full meaning of each presented sentence as early as possible, so that the model can learn to probabilistically preactivate the semantic features of all role fillers involved in the described event based on the statistical regularities in its environment.

Each training trial consists in randomly generating a new {*sentence, event*} pair based on the simple generative model depicted in Fig. 7a, and then going through the following steps: At the beginning of a sentence, all units are set to 0. Then, for each constituent of the sentence, the input unit or units representing the constituent are turned on and activation flows from the input units and – at the same time via recurrent connections - from the SG units to the units in the first hidden layer (Hidden 1), and from these to the units in the SG layer where the previous representation (initially all 0's) is replaced by a new activation pattern which reflects the influence of the current constituent. The activation pattern at the SG layer is then frozen while the model is probed concerning the event described by the sentence in the query part of the model. Specifically, for each probe question, a unit (representing a thematic role) or units (corresponding to feature-based representations of fillers concepts) at the probe layer are activated and feed into the hidden layer (Hidden 2) which at the same time receives activation from the SG layer. Activation from the SG and the probe layer combine and feed into the output layer where the units representing the complete role-filler pair (i.e., the unit representing the thematic role and the units corresponding to the feature-based representation of the filler concept) should be activated. After each presented constituent, the model is probed once for the filler of each role and once for the role of each filler involved in the described event, and for each response, the model's activation at the output layer is compared with the correct output. After each response, the gradient of the cross-entropy error measure for each connection weight and bias term in the query network is back-propagated through this part of the network, and the corresponding weights and biases are adjusted accordingly.

At the SG layer, the gradient of the cross-entropy error measure for each connection weight and bias term in the update network is collected for the responses on all the probes for each constituent before being back-propagated through this part of the network and adjusting the corresponding weights and biases. We used a learning rate of 0.00001 and momentum of 0.9 throughout.

Simulation of empirical findings. Because the model's implicit probabilistic representation of meaning and thus also the semantic update at any given point is determined by the statistical regularities in the training set, in the description of the simulations below we try to make clear how the observed effects depend on the training corpus (please refer to Fig. 7a).

For the simulations of semantic incongruity, cloze probability, and categorically related semantic incongruity, for each condition one agent ("man") was presented once with each of the ten specific actions (excluding only "like" and "look at"). The agent was not varied because the conditional probabilities for the later sentence constituents depend very little on the agents (the only effect of the choice of agent is that the manipulation of cloze probability is stronger for "man" and "woman", namely .7 vs. .3, than for "girl" and "boy", namely .6 vs. .4; see Fig. 7a). For the simulation of semantic incongruity, the objects were the high probability objects in the congruent condition (e.g., "The man plays *chess*.") and unrelated objects in the incongruent condition (e.g., "The man plays *salmon*"). For the simulation of cloze probability, the objects/patients were the high probability objects in the high cloze condition (e.g., "The man plays *chess*.") and the low probability objects in the low cloze condition (e.g., "The man plays *monopoly*."). For the simulation of categorically related semantic incongruities, the congruent and incongruent conditions from the semantic incongruity simulation were kept the same and there was an additional condition where the objects were from the same semantic category as the high and low probability objects related to the action (and thus shared semantic features at the output layer, e.g., "The man plays

backgammon"), but were never presented as patients of that specific action during training (so that their conditional probability to complete the presented sentence beginnings was 0). Instead, these objects only occurred as patients of the unspecific "like" and "look at" actions (Fig. 7a). For all these simulations, there were 10 items in each condition, and semantic update was computed based on the difference in SG layer activation between the presentation of the action (word *n-1*) and the object (word *n*).

For the simulation of the influence of a word's position in the sentence, we presented the longest possible sentences, i.e. all sentences that had occurred during training with a situation and a location, including both the version with the high probability ending and the version with the low probability ending of these sentences. There were 12 items in each condition, and semantic update was computed over the course of the sentences, i.e. the difference in SG layer activation between the first and the second word provided the basis for semantic update induced by the second word (the agent), the difference in SG layer activation between the second and the third word provided the basis for semantic update induced by the third word (the action), the difference in SG layer activation between the third and the fourth word provided the basis for semantic update induced by the fourth word (the object/ patient), and the difference in SG layer activation between the fourth and the fifth word provided the basis for semantic update induced by the fifth word (the location). It is interesting to consider the conditional probabilities of the constituents over the course of the sentence: Given a specific situation, the conditional probability of the presented agent ("man"; at the second position in the sentence) is .36 (because the conditional probability of that agent is overall .4, and the probability of the sentence being an active sentence such that the agent occurs in the second position is .9; see Fig. 7a). The conditional probability of the action (at the third position) is 1 because the actions are determined by the situations (see section on reversal anomalies, below, for the rationale behind this predictive relationship between the situation

and the action). The conditional probability of the objects (at the fourth position) is either .7 (for high probability objects) or .3 (for low probability objects) so that it is .5 on average, and the conditional probability of the location (at the fifth position) is 1 because the locations are determined by the objects. Thus, the constituents' conditional probabilities do not gradually decrease across the course of the sentences. The finding that semantic update nonetheless gradually decreased over successive words in these sentences (see *Results*) suggests that the SG layer activation does not perfectly track conditional probabilities. Even if an incoming word can be predicted with a probability of 1.0 so that an ideal observer could in principle have no residual uncertainty, the presentation of the item itself still produces some update, indicating that the model retains a degree of uncertainty, consistent with the 'noisy channel' model[47]. In this situation, as we should expect, the SG anticipates the presentation of the item more strongly as additional confirmatory evidence is accumulated, so that later perfectly predictable constituents are more strongly anticipated than earlier ones. In summary, the model's predictions reflect accumulation of predictive influences, rather than completely perfect instantaneous sensitivity to probabilistic constraints in the corpus.

For the simulation of lexical frequency, the high frequency condition comprised the high probability objects from the ten semantic categories, the two high probability agents ("woman" and "man") and two high probability locations ("kitchen" and "living room"). The low frequency condition contained the ten low probability objects, the two low probability agents ("girl" and "boy") and two low probability locations ("balcony" and "veranda"). The high and low frequency locations were matched pairwise in terms of the number and diversity of object patients they are related to ("kitchen" matched with "balcony", "living room" matched with "veranda"). Before presenting the high versus low frequency words, we presented a blank stimulus to the network (i.e., an input pattern consisting of all 0) to evoke the model's default activation which reflects the encoding of base-rate probabilities in the model's connection weights. There were 14 items in each condition, and semantic update was

computed based on the difference in SG layer activation between the blank stimulus (word *n-1*) and the high or low frequency word (word *n*).

To simulate semantic priming, for the condition of semantic relatedness, the low and high probability objects of each of the ten semantic object categories were presented subsequently as prime-target pair (e.g., "monopoly chess"). For the unrelated condition, primes and targets from the related pairs were re-assigned such that there was no semantic relationship between prime and target (e.g., "sunfish chess"). For the simulation of associative priming, the condition of associative relatedness consisted of the ten specific actions as primes followed by their high probability patients as targets (e.g., "play chess"). For the unrelated condition, primes and targets were again re-assigned such there was no relationship between prime and target (e.g., "play eggs"). To simulate repetition priming, the high probability object of each semantic category was presented twice (e.g., "chess chess"). For the unrelated condition, instead of the same object, a high probability object from another semantic category was presented as prime. For all priming simulations, there were 10 items in each condition, and semantic update was computed based on the difference in SG layer activation between the prime (word *n-1*) and the target (word *n*).

For the simulation of semantic illusions/ reversal anomalies, each of the eight situations was presented, followed by the high probability object related to that situation and the action typically performed in that situation (e.g., "At breakfast, the eggs *eat*…"). For the congruent condition, the situations were presented with a possible agent and the action typically performed in that situation (e.g., "At breakfast, the man *eats*…") and for the incongruent condition, with a possible agent and an unrelated action (e.g., "At breakfast, the man *plants*…"). There were eight items in each condition, and semantic update was computed based on the difference in SG layer activation between the presentation of the second constituent which could be an object or an agent (e.g., "eggs" or "man"; word *n-1*) and the action (word *n*). Please note that in the model environment, the situations predict specific

actions with a probability of 1. This prevented the critical words (i.e., the actions) from being much better predictable in the reversal anomaly condition where they are preceded by objects (which in the model environment also predict specific actions with a probability of 1) as compared to the congruent condition where they are preceded by agents (which are not predictive of specific actions at all). Of course, situations do not completely determine actions in the real world. However, the rationale behind the decision to construct the corpus in that way to simulate the reversal anomaly experiment by Kuperberg and colleagues[19] was that the range of plausibly related actions might be similar for specific situations and specific objects such that actions are not much better predictable in the reversal anomaly than in the congruent condition. A relevant difference between both conditions was that in the reversal anomaly condition the model initially assumed the sentences to be in passive voice, because during training, sentences with the objects presented before the actions had always been in passive voice (see Fig. 7a). Thus, when the critical word was presented without passive marker (i.e., "by"), the model revised its initial assumptions in that regard in the reversal anomaly condition while there was no need for revision in the congruent condition.

To simulate the developmental trajectory of N400 effects we examined the effect of semantic incongruity on semantic update (as described above) at different points in training, specifically after exposure to 10000, 100000, 200000, 400000, and 800000 sentences. To examine the relation between update at the SG layer and update at the output layer (reflecting latent and explicit estimates of semantic feature probabilities, respectively), at each of the different points in training (see above) we computed the update of activation at the output layer (summed over all role filler pairs) analogously to the activation update at the SG layer.

To simulate semantic priming effects on N400 amplitudes during near-chance lexical decision performance in a second language, we examined the model early in training when it had been presented with just 10000 sentences. As illustrated in Figure 5a, at this point the model fails to understand words and sentences, i.e. to activate the corresponding units at the

output layer. The only knowledge that is apparent in the model's performance at the output layer concerns the possible filler concepts for the agent role and their relative frequency, as well as a beginning tendency to activate the correct agent slightly more than the others. Given the high base-rate frequencies of the possible agents, it does not seem surprising that the model learns this aspect of its environment first. At this stage in training, we simulated semantic priming as described above. In addition, even though this has not been done in the empirical study, we also simulated associative priming and influences of semantic incongruity in sentences (as described above).

For the simulation of the interaction between semantic incongruity and repetition, all sentences from the simulation of semantic incongruity (see above) were presented twice, in two successive blocks (i.e., running through the first presentation of all the sentences before running through the second presentation) with connection weights being adapted during the first round of presentations (learning rate = .01). Sentences were presented in a different random order for each model with the restrictions that the presentation order was the same in the first and the second block, and that the incongruent and congruent version of each sentence directly followed each other. The order of conditions, i.e. whether the incongruent or the congruent version of each sentence was presented first was counterbalanced across models and items (i.e., for half of the models, the incongruent version was presented first for half of the items, and for the other half of the models, the incongruent version was presented first for the other half of the items).

It is often assumed that learning is based on prediction error[27–29]. Because the SG layer activation at any given time represents the model's implicit prediction or probability estimates of the semantic features of all aspects of the event described by a sentence, the change in activation induced by the next incoming word can be seen as the prediction error contained in the previous representation (at least as far as it is revealed by that next word). Thus, in accordance with the widely shared view that prediction errors drive learning, we used a

temporal difference (TD) learning approach, assuming that in the absence of observed events, learning is driven by this prediction error concerning the next internal state. Thus, the SG layer activation at the next word serves as the target for the SG layer activation at the current word, so that the error signal becomes the difference in activation between both words, i.e. $SG_{n+1} - SG_n$ (also see section *Semantic update driven learning rule*, above). There were 10 items in each condition, and semantic update was computed during the first and second presentation of each sentence as the difference in SG layer activation between the presentation of the action (word *n*-1) and the object (word *n*).

Statistics

All reported statistical results are based on ten runs of the model each initialized independently (with initial weights randomly varying between +/- .05) and trained with independently-generated training examples as described in section *Simulation Details/ Environment* (N=800000, unless otherwise indicated). In analogy to subject and item analyses in empirical experiments, we performed two types of analyses on each comparison, a model analysis with values averaged over items within each condition and the 10 models treated as random factor, and an item analysis with values averaged over models and the items (N ranging between 8 and 14; please see the previous section for the exact number of items in each simulation experiment) treated as random factor. There was no blinding. We used two-sided paired t-tests to analyze differences between conditions; when a simulation experiment involved more than one comparison, significance levels were Bonferroni-corrected within the simulation experiment. To test for the interaction between repetition and congruity, we used a repeated measures analysis of variance (rmANOVA) with factors Repetition and Congruity. To analyze whether our data met the normality assumption for these parametric tests, we tested differences between conditions (for the t-tests) and residuals (for the rmANOVA) for normality with the Shapiro-Wilk test. Using study-wide Bonferroni correction to adjust significance levels for the multiple performed tests, results did not show significant deviations

from normality (all *p*s > .11 for the model analyses and > .25 for the item analyses). However, to further corroborate our results we additionally tested all comparisons with deviations from normality at uncorrected significance levels <.05 using the Wilcoxon signed rank test which does not depend on the normality assumptions; all results remained significant. Specifically, in the model analyses deviations from normality at uncorrected significance levels were detected for the semantic incongruity effect (Fig. 2a; $p = .043$) and the frequency effect (Fig. 2e; $p = .044$), as well as for the difference between categorically related incongruities and congruent completions (Fig. 2d; $p = .0053$). Wilcoxon signed rank tests confirmed significant effects of semantic incongruity (Fig. 2a; $p = .002$) and lexical frequency (Fig. 2e; $p = .037$), and a significant difference between categorically related incongruities and congruent sentence continuations (Fig. 2d; $p = .002$). In the item analyses, deviations from normality at an uncorrected significance level were detected for the difference between incongruent completions and semantic illusions (Supplementary Fig. 1i; $p = .012$). Again, the Wilcoxon signed rank test confirmed a significant difference between the conditions ($p = .0078$).

Using Levene's test, we detected violations of the assumption of homogeneity of variances (required for the rmANOVA used to analyze the interaction between repetition and congruity; Fig. 6 and Supplementary Fig. 4) in the item analysis, $F_2(3) = 12.05$, $p < .0001$, but not in the model analysis, $F_1 < 1$. We nonetheless report the ANOVA results for both analyses because ANOVAs are typically robust to violations of this assumption as long as the groups to be compared are of the same size. However, we additionally corroborated the interaction result from the item ANOVA by performing a two-tailed paired t-test on the repetition effects in the incongruent versus congruent conditions, i.e. we directly tested the hypothesis that the size of the difference in the model's N400 correlate between the first presentation and the repetition was larger for incongruent than for congruent sentence completions: incongruent (first – repetition) > congruent (first – repetition). Indeed, the size of the repetition effects significantly differed between congruent and incongruent conditions, $t_{2(9)} = 10.99$, $p < .0001$,

and the differences between conditions did not significantly deviate from normality, $p = .44$, thus fulfilling the prerequisites for performing the t-test.

In general, systematic deviations from normality are unlikely for the results by-model (where apparent idiosyncrasies are most probably due to sampling noise), but possible in the by-item data. Thus, while we present data averaged over items in the figures in the main text in accordance with the common practice in ERP research to analyze data averaged over items, for transparency we additionally display the data averaged over models as used for the by-item analyses (see Supplementary Fig. 1-4).

Code availability

All computer code used to run the simulations and analyze the results will be made available on github at the time of publication.

Author contributions

M.R. developed the idea for the project, including the idea of linking the N400 to the updating of SG layer activation in the model. S.S.H. re-implemented the model for the current simulations. M.R. and J.L.M. formulated the training environment. J.L.M. formulated the new learning rule and developed the probabilistic formulation of the model with input from M.R. M.R. adjusted the model implementation, implemented the training environment, formulated and implemented the simulations, trained the networks and conducted the simulations, and performed the analyses with input from J.L.M. J.L.M. and M.R. discussed the results and wrote the manuscript.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 658999 to Milena Rabovsky. We thank Roger Levy, Stefan Frank, and the members of the PDP lab at Stanford for helpful discussion.

References

1. Kutas, M. & Hillyard, S. A. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science (80-.).* **207,** 203–205 (1980).
2. Kutas, M. & Federmeier, K. D. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62,** 621–647 (2011).
3. Lau, E. F., Phillips, C. & Poeppel, D. A cortical network for semantics: (de)constructing the N400. *Nat. Rev. Neurosci.* **9,** 920–933 (2008).
4. Brown, C. & Hagoort, P. The processing nature of the N400: Evidence from masked priming. *J. Cogn. Neurosci.* **5,** 34–44 (1993).
5. Hagoort, P., Baggio, G. & Willems, R. M. in *The Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 819–836 (MIT Press, 2009).
6. Federmeier, K. D. & Laszlo, S. *Chapter 1 Time for Meaning. Electrophysiology Provides Insights into the Dynamics of Representation and Processing in Semantic Memory*. *Psychology of Learning and Motivation - Advances in Research and Theory* **51,** (2009).
7. Debruille, J. B. The N400 potential could index a semantic inhibition. *Brain Res. Rev.* **56,** 472–477 (2007).
8. McClelland, J. L., St. John, M. & Taraban, R. Sentence comprehension: A parallel distributed processing approach. *Lang. Cogn. Process.* **4,** 287–336 (1989).
9. Itti, L. & Baldi, P. Bayesian Surprise Attracts Human Attention. 1–8 (2006). doi:10.1016/j.visres.2008.09.007
10. Kutas, M. & Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature* **307,** 101–103 (1984).
11. Van Petten, C. & Kutas, M. Influences of semantic and syntactic context on open- and closed-class words. *Mem. Cogn.* **19,** 95–112 (1991).
12. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106,** 1126–1177 (2008).
13. Frank, S. L., Galli, G. & Vigliocco, G. The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* **140,** 1–25 (2015).
14. Federmeier, K. D. & Kutas, M. A Rose by Any Other Name : Long-Term Memory Structure and Sentence Processing. *J. Mem. Lang.* **41,** 469–495 (1999).
15. Barber, H., Vergara, M. & Carreiras, M. Syllable-frequency effects in visual word recognition: evidence from ERPs. *Neuroreport* **15,** 545–548 (2004).
16. Koivisto, M. & Revonsuo, A. Cognitive representations underlying the N400 priming effect. *Cogn. Brain Res.* **12,** 487–490 (2001).
17. Rugg, M. D. The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology* **22,** 642–647 (1985).
18. Laszlo, S. & Plaut, D. C. A neurally plausible Parallel Distributed Processing model of Event-Related Potential word reading data. *Brain Lang.* **120,** 271–281 (2012).

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

20. Brouwer, H., Crocker, M. W., Venhuizen, N. j & Hoeks, J. C. J. A Neurocomputational Model of the N400 and the P600 in Language Comprehension. *Cogn. Sci.* (in press).

21. Kim, A. & Osterhout, L. The independence of combinatory semantic processing: Evidence from event-related potentials. *J. Mem. Lang.* **52,** 205–225 (2005).

22. Friedrich, M. & Friederici, A. D. N400-like semantic incongruity effect in 19-month-olds: Processing known words in picture contexts. *J. Cogn. Neurosci.* **16,** 1465–77 (2004).

23. Atchley, R. A. *et al.* A comparison of semantic and syntactic event related potentials generated by children and adults. *Brain Lang.* **99,** 236–246 (2006).

24. Kutas, M. & Iragui, V. The N400 in a semantic categorization task across 6 decades. *Electroencephalogr. Clin. Neurophysiol. - Evoked Potentials* **108,** 456–471 (1998).

25. Gotts, S. J. Incremental learning of perceptual and conceptual representations and the puzzle of neural repetition suppression. (2015). doi:10.3758/s13423-015-0855-y

26. McLaughlin, J., Osterhout, L. & Kim, A. Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nat. Neurosci.* **7,** 703–704 (2004).

27. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science (80-.).* **275,** 1593–1599 (1997).

28. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **360,** 815–36 (2005).

29. McClelland, J. L. The interaction of nature and nurture in development: A parallel distributed processing perspective. *Int. Perspect. Psychol. Sci. Vol. 1 Lead. Themes* (1994).

30. Besson, M., Kutas, M. & Petten, C. Van. An Event-Related Potential (ERP) Analysis of Semantic Congruity and Repetition Effects in Sentences. *J. Cogn. Neurosci.* **4,** 132–149 (1992).

31. Schott, B., Richardson-Klavehn, A., Heinze, H.-J. & Düzel, E. Perceptual priming versus explicit memory: dissociable neural correlates at encoding. *J. Cogn. Neurosci.* **14,** 578–592 (2002).

32. Laszlo, S. & Armstrong, B. C. PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended Event-Related Potential reading data. *Brain Lang.* **132,** 22–27 (2014).

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

34. Friederici, A. D. Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* **6,** 78–84 (2002).

35. Coulson, S., King, J. W. & Kutas, M. ERPs and Domain Specificity: Beating a Straw Horse. *Lang. Cogn. Process.* **13,** 653–672 (1998).

36. McCandliss, B. D., Cohen, L. & Dehaene, S. The visual word form area: Expertise for reading in the fusiform gyrus. *Trends Cogn. Sci.* **7,** 293–299 (2003).

37. Rumelhart, D. E. in *Metaphor and Thought* (ed. Ortony, A.) 71–82 (Cambridge University Press, 1979).

38. St. John, M. F. & McClelland, J. L. Learning and applying contextual constraints in sentence comprehension. *Artif. Intell.* **46,** 217–257 (1990).

39. Rohde, D. L. T. A Connectionist Model of Sentence Comprehension and Production. (Carnegie Mellon University, 2002).

40. Bryant, B. D. & Miikkulainen, R. *From Word Stream to Gestalt: A Direct Semantic Parse for Complex Sentences*. (2001).

41. Hinton, G. I. Connectionist Learning Procedures. *Mach. Learn. -- an Artif. Intell. Approach* **III,** 555–610 (1990).

42. Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. in *Parallel Distributed Processing* (eds. Rumelhart, D. E. & McClelland, J. L.) 77–109 (MIT Press, 1986). doi:10.1146/annurev-psych-120710-100344

43. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).

44. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. *Emnlp2014.Org* at <http://emnlp2014.org/papers/pdf/EMNLP2014162.pdf>

45. Rumelhart, D. E. & Todd, P. M. Learning and connectionist representations. *Atten. Perform. XIV Synerg. Exp. Psychol. Artif. Intell. Cogn. Neurosci.* 3–30 (1993).

46. McClelland, J. L. & Rogers, T. T. The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* **4,** 310–322 (2003).

47. Levy, R., Bicknell, K., Slattery, T. & Rayner, K. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proc. Natl. Acad. Sci.* **106,** 21086–21090 (2009).

Simulated effects	Example	N400 data	Reference
Basic effects			
Semantic incongruity	I take my coffee with cream and *sugar/ dog.*	cong. < incong.	Kutas & Hillyard (1980)
Cloze probability	Don't touch the wet *paint/ dog.*	high < low	Kutas & Hillyard (1984)
Position in sentence		late < early	Van Petten & Kutas (1991)
Categorically related incongruity	They wanted to make the hotel look more like a tropical resort. So along the driveway they planted rows of *palms/ pines/ tulips*.	cong. < cat. rel. incong. < incong.	Federmeier & Kutas (1999)
Lexical frequency		high < low	Barber, Vergara, & Carreiras (2004)
Semantic priming	sofa - bed	related < unrelated	Koivisto & Revonsuo (2001)
Associative priming	wind - mill	related < unrelated	Koivisto & Revonsuo (2001)
Repetition priming		repeated < unrelated	Rugg (1985)
Reversal anomalies	Every morning at breakfast the boys would *eat/* Every morning at breakfast the eggs would *eat/* Every morning at breakfast the boys would *plant*	cong. =< rev. anom. < incong.	Kuperberg, Sitnikova, Caplan, & Holcomb (2003)
Extensions			
Age		babies: less compr. < more compr. later: young > old	Friedrich & Friederici (2009), Kutas & Iragui (1998), Atchley et al. (2006)
Priming during near chance 2nd language performance	chien – chat	related < unrelated	McLaughlin, Osterhout & Kim (2004)
Repetition X incongruity		cong. (|nonrep. – rep.|) < incong. (|nonrep. – rep.|)	Besson, Kutas, & van Petten (1992)

Table 1. Overview of simulated effects. cong: congruent; incong.: incongruent; cat. rel.: categorically related; rev. anom.: reversal anomaly; compr.: comprehension; rep.: repeated; nonrep.: nonrepeated.

Supplementary Table 1

Words (i.e. labels of input units) and their semantic representations (i.e., labels of the output units by which the concepts that the words refer to are represented)

Words	Semantic representations
Woman	person, agent, adult, female, woman
Man	person, agent, adult, male, man
Girl	person, agent, child, female, girl
Boy	person, agent, child, male, boy
Drink	action, consume, done with liquids, drink
Eat	action, consume, done with foods, eat
Feed	action, done to animals, done with food, feed
Fish	action, done to fishes, done close to water, fish
Plant	action, done to plants, done with earth, plant
Water	action, done to plants, done with water, water
Play	action, done with games, done for fun, play
Wear	action, done with clothes, done for warming, wear
Read	action, done with letters, perceptual, read
Write	action, done with letters, productive, write
Look at	action, visual look at
Like	action, positive, like
Kitchen	location, inside, place to eat, kitchen
Living room	location, inside, place for leisure, living room
Bedroom	location, inside, place to sleep, bedroom
Garden	location, outside, place for leisure, garden
Lake	location, outside, place with animals, lake
Park	location, outside, place with animals, park
Balcony	location, outside, place to step out, balcony
River	location, outside, place with water, river
Backyard	location, outside, place behind house, backyard
Veranda	location, outside, place in front of house, veranda
Breakfast	situation, food related, in the morning, breakfast
Dinner	situation, food related, in the evening, dinner
Excursion	situation, going somewhere, to enjoy, excursion
Afternoon	situation, after lunch, day time, afternoon
Holiday	situation, special day, no work, holiday
Sunday	situation, free time, to relax, Sunday
Morning	situation, early, wake up, morning
Evening	situation, late, get tired, evening
Egg	consumable, food, white, egg
Toast	consumable, food, brown, toast
Cereals	consumable, food, healthy, cereals
Soup	consumable, food, in bowl, soup

Pizza	consumable, food, round, pizza
Salad	consumable, food, light, salad
Iced tea	consumable, drink, from leaves, iced tea
Juice	consumable, drink, from fruit, juice
Lemonade	consumable, drink, sweet, lemonade
Cacao	consumable, drink, with chocolate, cacao
Tea	consumable, drink, hot, tea
Coffee	consumable, drink, activating, coffee
Chess	game, entertaining, strategic, chess
Monopoly	game, entertaining, with dice, monopoly
Backgammon	game, entertaining, old, backgammon
Jeans	garment, to cover body, for legs, jeans
Shirt	garment, to cover body, for upper part, shirt
Pajamas	garment, to cover body, for night, pajamas
Novel	contains language, contains letters, art, novel
Email	contains language, contains letters, communication, email
SMS	contains language, contains letters, communication, short, SMS
Letter	contains language, contains letters, communication, on paper, letter
Paper	contains language, contains letters, scientific, paper
Newspaper	contains language, contains letters, information, newspaper
Rose	can grow, has roots, has petals, red, rose
Daisy	can grow, has roots, has petals, yellow, daisy
Tulip	can grow, has roots, has petals, colorful, tulip
Pine	can grow, has roots, has bark, green, pine
Oak	can grow, has roots, has bark, tall, oak
Birch	can grow, has roots, has bark, white bark, birch
Robin	can grow, can move, can fly, red, robin
Canary	can grow, can move, can fly, yellow, canary
Sparrow	can grow, can move, can fly, brown, sparrow
Sunfish	can grow, can move, can swim, yellow, sunfish
Salmon	can grow, can move, can swim, red, salmon
Eel	can grow, can move, can swim, long, eel
By	passive voice (activated together with the deep subject, e.g., ‘by the man’)
Was	passive voice (activated together with the verb, e.g., ‘was played’)
During/at	no output units (activated together with situation words, e.g., ‘at breakfast’)
In	no output units (activated together with location words, e.g., ‘in the park’)