I like coffee with cream and *dog*? Change in an implicit probabilistic representation captures meaning processing in the brain

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

Department of Psychology, Stanford University

Word count: 8436

*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. This component, however, has been hard to capture within traditional approaches to language processing. Here, we show that a neural network that eschews these traditions can capture a wide range of findings on the factors that affect the amplitude of the N400. The model simulates the N400 as the change induced by an incoming word in an initial, implicit probabilistic representation of the situation or event described by the linguistic input, captured by the hidden unit activation pattern in a neural network. We further propose a new learning rule in which the process underlying the N400 drives implicit learning in the network. The model provides a unified account of a large body of findings and connects human language processing with successful deep learning approaches to language processing.

I like coffee with cream and *dog*? Change in an implicit probabilistic representation captures meaning processing in the brain

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 basis of meaning processing. The N400 is a negative deflection at centroparietal electrode sites peaking around 400 ms after the presentation of a potentially meaningful stimulus. The first report of the N400 showed that it occurred on presentation of a word violating expectations established by 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 range of variables including sentence context, category membership, 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 are actively debated[3–7], but their capacity to capture all the relevant data is difficult to stringently evaluate due to the lack of implementation and none has yet offered a generally accepted account. Indeed, the authors of a recent review (Kutas & Federmeier, 2011) have noted that "ERP parameters are (…) neither generally nor readily reducible to psychological constructs." Ultimately, the authors suggest "the field must be willing to rethink the pool of available cognitive constructs it has developed, largely from end-state measures" (p. 624). Existing accounts are often grounded, at least in part, in traditional modes of theorizing based on constructs originating in the 1950's[8], in which symbolic representations (e.g., of the meanings of words) are retrieved from memory and subsequently integrated into a compositional representation – an annotated structural description thought to serve as the representation of the meaning of a sentence[9–11]. Even though perspectives on language processing have evolved in a variety of ways, many researchers maintain the notion that word meanings are first retrieved from memory and

subsequently assigned to roles in a compositional representation. The account we offer here does not employ these constructs and thus may contribute to the effort to rethink aspects of several foundational issues: What does it mean to understand language? What are the component parts to the process? Do we construct a structural description of a spoken utterance in our mind, or do we more directly construct a representation of the speaker's meaning? Our work suggests different answers than those often given to these questions.

We present an explicit computational model that accounts well for a wide range of findings in the literature on the N400. The model, called the *Sentence Gestalt* (SG) model, was initially developed nearly 30 years ago [12,13] with the explicit goal of illustrating how language understanding might occur *without* relying on the traditional mode of theorizing described above. The model sought to offer a functional-level characterization of language understanding in which each word in a sentence someone hears or reads provides clues that constrain the formation of an implicit representation of the event being described by the sentence. The initial work with the model[13] established that it could capture several core aspects of language, including the ability to resolve ambiguities of several kinds; to use word order and semantic constraints in constructing the event representation; and to represent events described by sentences never seen during the network's training.

The current work extending this model to address N400 amplitudes complements efforts to model neurophysiological details underlying the N400[14–16]; we focus on providing a functional level account of the way the probabilistic relationship between linguistic utterances and their meanings – and human experience of this relationship – shapes the extent to which the presentation of a word or sequence of words updates a learned representation of meaning, defined as an implicit representation that supports accurate estimates of the probability of the different aspects of the event described by the sentence.

The design of the model reflects the principle that listeners continually update their representation of the event being described as each incoming word of a sentence is presented.

The representation is an internal representation (putatively 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 aspects or features of the event being described by the sentence and of the participants in the event (see *Implicit probabilistic theory of 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. Specifically, the *semantic update* (SU) induced by the current word *n* is defined as the sum across the units in the SG layer of the absolute value of the change in each unit's activation produced by the current word *n*. For a given unit (indexed below by the subscript *i*), the change is simply the difference between 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[17] and to the signals that govern learning in the network (see *online methods* and below). Indeed, we propose a new learning rule driven by the semantic update, allowing the model to address how language processing even in the absence of external event information can drive learning about events and about how speakers use language to describe them.

How does the semantic update capture the N400? After a listener has heard "I take my coffee with cream and…" our account holds that the activation state already implicitly represents a high subjective probability 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 change in activation, and thus a small N400 amplitude. In contrast, the representation will change much more if "…dog" is presented instead, corresponding to a

much larger change in the subjective probability, reflected in a larger change in the pattern of activation and thus a larger N400 amplitude.

Distinctive Features of the Sentence Gestalt Model

Several aspects of the model's design and behavior are worth understanding in order to see why it accounts for the findings we apply it to below. First, the model is designed to form a representation of the *event* described by the sentence that it hears, rather than a representation of the sentence itself. The words (and their arrangement) provide *clues* about the event, and objects can be inferred as event participants without being mentioned. For example, a knife might be inferred upon hearing 'The boy spread the butter on the bread.' This makes it different from many other models of language processing in which listeners are thought to be updating specifically *linguistic* expectations about specific words or to be building structured representations in which word meanings are inserted into roles or slots in a structural description that may be tied closely to the sentence itself[10,11]. Furthermore, unlike most other models, the SG model does not contain separate modules that implement distinct stages of syntactic parsing or of accessing the meanings of individual words on the way to the formation of a representation of the event. Instead the model simply maps from word forms to an implicit probabilistic representation of the overall meaning of the sentence.

Second, we as modelers make no stipulations of the form or structure of the model's internal representations[1]. Rather, these representations are shaped by the statistics of the experiences it is trained on, as in some language representation models developed by other groups in recent years [18,19]. In this way our model is similar to contemporary deep learning models such as Google Translate[20], which likewise make no stipulations of the form or structure of the internal representation generated from an input sentence; instead the

[1] To train the model, the model does require a way of providing it with information about the event described by the sentence. We follow the choice made in the original implementation, in which events are described in terms of an action, a location, a situation (such as 'at breakfast'), the actor or agent in the event, and the object or patient to which the action is applied. Critically, the event description is not the model's internal representation of the event, but is instead a simplified characterization of those aspects of events that the model learns to derive from the presented sentences.

representations are shaped by the process of learning to predict the translation of an input sentence in one language into other languages. Though our model is simpler than Google Translate, which employs more layers of neuron-like processing units, the models are similar in avoiding representational commitments, and the success of Google Translate can be seen as supporting the view that a commitment to *any* stipulated form of internal representation is an impediment to capturing the nuanced, *quasiregular* nature of language[21,22]. Learning takes place in the model over an extended time course thought of as loosely corresponding to the time course of human development into early adulthood, based on the gradual accumulation of experience about events and the sentences speakers use to describe them. Among other things, this means that the extent of the semantic update that occurs upon the presentation of a particular word in a particular context depends not only on the statistics of the environment, but also on the extent of the model's training – thereby allowing it to address changes in N400 responses as a function of experience.

Third, the model responds to whatever inputs it receives, independently of whether its inputs form sentences or are simply isolated words or pairs of words. Thus the model will update its state after the presentation of any word, allowing the possibility of capturing findings from N400 studies in which words are presented singly or in pairs, as well as findings from studies in which N400's are observed to words presented in complete sentence contexts.

Finally, we view the processing of language (and other forms of meaningful input) to be a complex and multi-faceted process, and we see the SG model – and the N400 – as characterizing one aspect of this process. This view is consistent with the fact that other ERP components appear to reflect different aspects of language processing. Specifically, we see the model as reflecting an implicit process that operates quickly and automatically as a stream of linguistic input is presented, constructing an implicit, initial representation of the event or situation that is being described. Language processing may also involve other components

that might form expectations about specific word-forms and their sequencing that are not captured by the SG model or the N400. Furthermore, the initial representation that the model forms as it processes language in real time may not always correspond to the final understood meaning of a sentence. Other processes may come into play in understanding sentences with unusual structure, and these processes may result in changes to the meaning representation that is ultimately derived from reading or listening to a linguistic input. In the *Discussion* below we consider how the formation of an initial, implicit representation of meaning, as captured by the SG model, might fit into this broader picture, and how our findings may inform discussions of other aspects of human language processing.

Training the Sentence Gestalt Model

To train the model, we use an artificial corpus of {*sentence*, *event*} training examples produced by a generative model that embodies a simplified and controlled micro-world in which the statistics of events, the properties of the objects that occur in them, and the words used in sentences about these events are completely controlled by the modeler (see *online methods*). This approach prevents us from testing the model with the actual sentences used in targeted experiments, since the true statistics of real events and sentences are not fully captured. Given the successes of Google Translate and other deep learning approaches to language processing, it may eventually be possible to train a successor to our model on a much larger corpus of real sentences, allowing modeling of the semantic update produced by the actual materials used in empirical experiments. Such a success would still leave open the question of what factors were responsible for the model's behavior. Our approach, relying on a synthetic corpus, allows us to build into the training materials manipulations of variables corresponding to those explored in the designs of the experiments we are modeling. For example, we can separately manipulate how frequently an object designated by a particular word appears in an event of a particular type (e.g. how often a knife is used for spreading butter on bread) and the extent to which the properties of the object signaled by a word are

consistent with the properties of the objects that typically appear in events of this type (e.g. an axe, though never used in spreading, is more semantically similar to a knife than a chair is). Thus we are able to separate predictability from semantic similarity more cleanly than might be possible using a large corpus of real sentences.

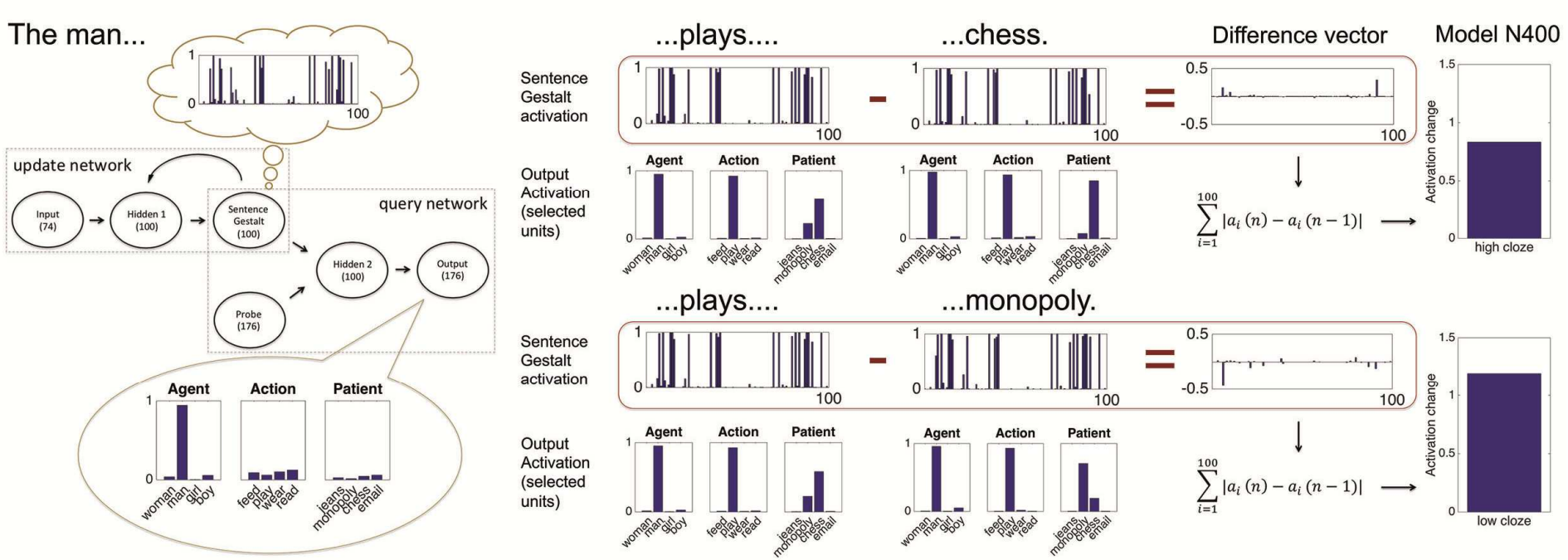


Figure 1. *The Sentence Gestalt (SG) model architecture, shown processing a sentence with a high or low cloze probability ending, and the model's N400 correlate. The model (gray boxes on the left) consists of an update network and a query network. Ovals represent layers of units (and number of units 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, each incoming word is processed through layer Hidden 1 where it combines with the previous activation of the SG layer to produce the updated SG pattern corresponding to the updated implicit representation of the event described by 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 layer Hidden 2 with the current SG pattern to produce output activations. Output units for selected query response units activated in response to the agent, action, and patient probes are shown; each query response includes a distinguishing event feature (e.g. 'man', 'woman', as shown) as well as other features (e.g., 'person', 'adult', not shown) that capture semantic similarities among event participants; see Supplementary Table 1). After presentation of "The man", the SG representation (thought bubble at top left) supports activation of the correct event features when probed for the agent and estimates the probabilities of action and patient features consistent with this agent. After the word "plays" (shown twice in the middle of the figure) the SG representation is updated and the model now activates the correct features given the agent and action probes, and estimates the probability of alternative possible patients. These estimates reflect the model's experience, since the man plays chess with higher probability than monopoly. If the next word is "chess" (top), the change in the pattern of activation on the SG layer (summed magnitudes of changes shown in 'Difference vector') is smaller than if the next word is "monopoly" (bottom). The change signal, called the Semantic Update (SU) is the proposed N400 correlate (right). It is larger for the less probable ending (monopoly, bottom) as compared to the more probable ending (chess, top).*

Results

We report fourteen simulations of well-established N400 effects chosen to illustrate how the model can address a broad range of empirical findings taken as supporting diverse and sometimes conflicting descriptive theories of the functional basis of the N400 (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). 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)"[23]. 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[24]. 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).

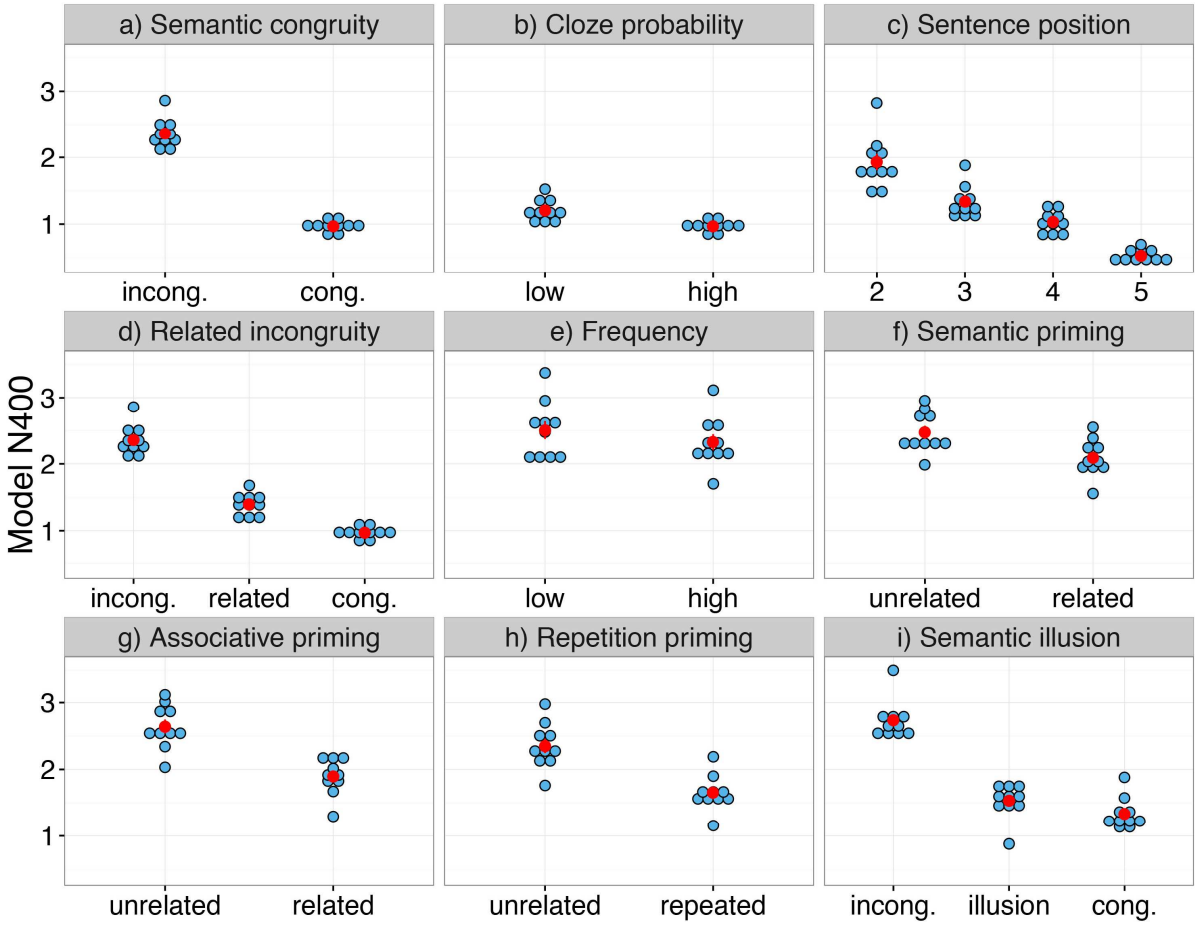
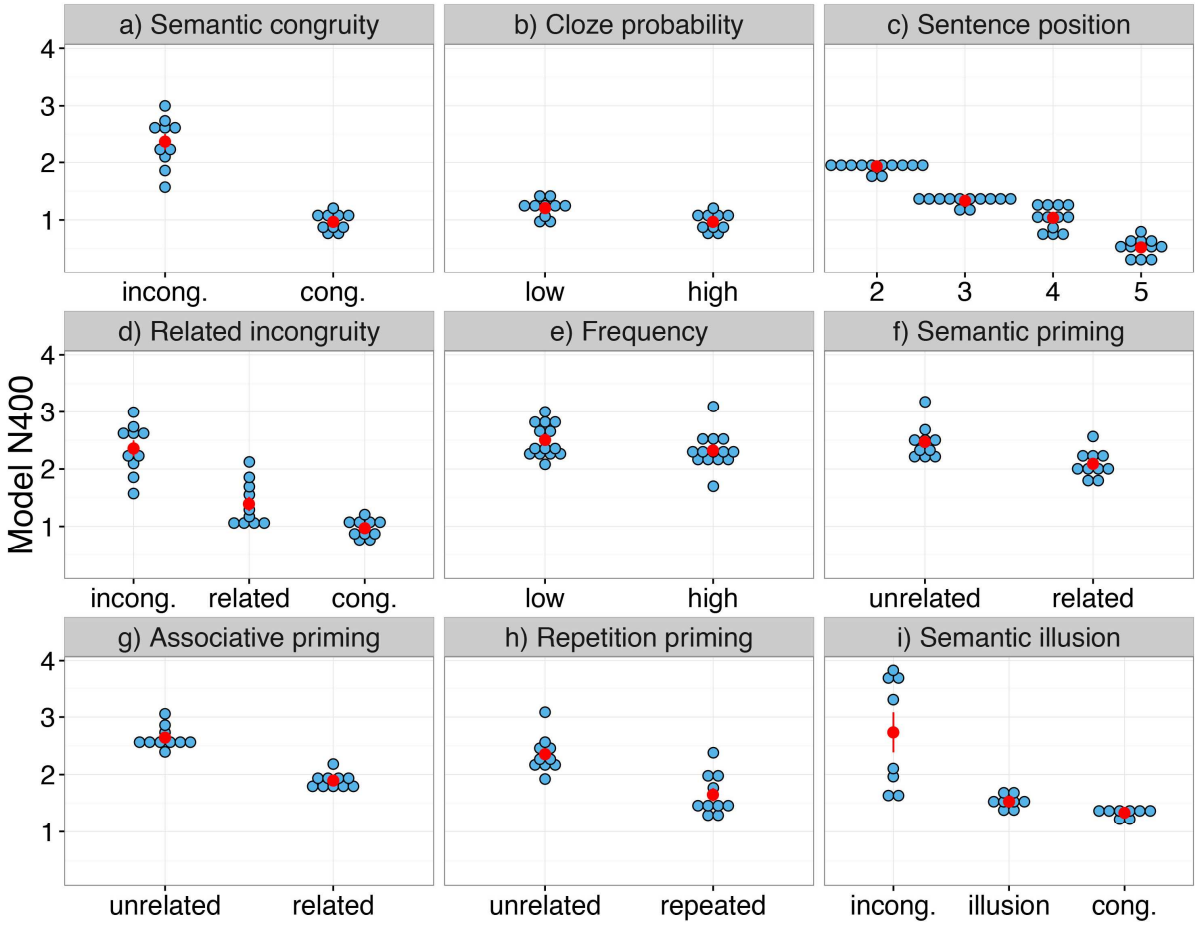


Figure 2. *Simulation results for the basic effects. Displayed is the model's N400 correlate, i.e. the update of the Sentence Gestalt layer activation – the model's probabilistic representation of sentence meaning - induced by the new incoming word. Cong., congruent; incong., incongruent. See text for 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.*

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[25]), 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[26]. 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 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)"[27]. 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 that shared semantic features capturing basic level category membership as well as other aspects of semantic similarity with high probability completions compared to those that did not share semantic features with any of the 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[7,28]. 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[29]; 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[30]; and for an immediate repetition of a word compared to the same word following an unrelated prime[31]. 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[3,14]. As previously noted, the SG pattern

probabilistically represents the meaning of a sentence if one is presented, but the model will 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*..."[32]. 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*..."[32]. How can we understand this pattern? One analysis[33] 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 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[34]. Since in the model's experience eggs are things that are eaten, and never things that eat, it continues to treat them this way even though the sentence structure differs from the structure it has experienced during training. To demonstrate the robustness of this kind of behavior in the model, we conducted an additional simulation of a similar finding using a slightly different type of reversal anomaly that has been the focus of a previous model[33] (see *discussion* section for more details). The experiment was conducted in Dutch using Dutch word order conventions, and differed from the previous study in that two noun phrases are presented prior to the presentation of the verb. In the anomalous sentences, the sentences seem to describe impossible events such as for instance an event in which a javelin throws some athletes (e.g. "De speer heft de atleten *geworpen*", lit: "The javelin has the athletes *thrown*"), yet there is little or no N400 response at the presentation of the verb relative to ordinary control sentences in which it is the athletes that are said to do the throwing ("De speer werd door de atleten *geworpen*", lit: "The javelin was

by the athletes *thrown*") [35]. For this simulation we trained an additional model on the same training corpus, but with Dutch word order (please see *online methods* and Supplementary Fig. 2 for details). Once again the SG model is not 'thrown' by the anomalous sentences – instead it interprets both versions of the sentences in a way that is consistent with its experience with events.

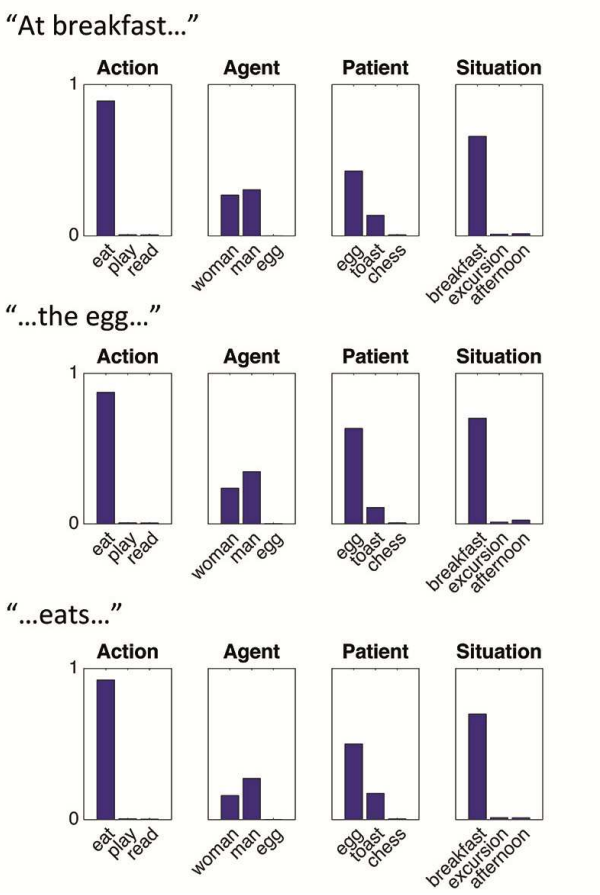
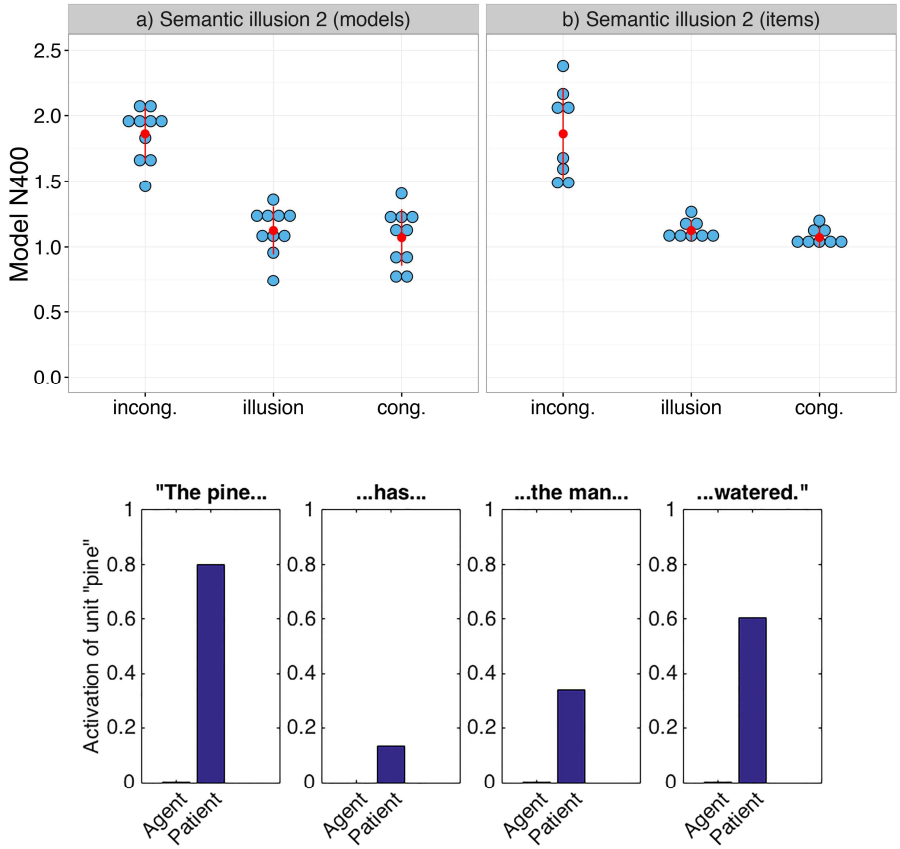


Figure 3. *Processing semantic illusions. Activation of selected output units while the model processes a sentence from the semantic illusion simulation: "At breakfast, the egg eats…". Note that the model continues to represent the egg as the patient (not the agent) of eating, even after the word "eat" 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. As noted in the introduction, our account leaves open the possibility that other processes which may be reflected in the P600 could be involved in detecting the anomaly and possibly revising the initial interpretation captured by the SG model (see *Discussion* below).



Supplementary Figure 2. Simulation results for a second type of semantic illusion where the relationship between two noun phrases is established prior to encountering the verb (see text for more details)[35]*; the simulation was conducted with a model trained with Dutch word order. Cong., congruent; incong., incongruent. Top left. Each blue dot represents the results for one independent run of the model, averaged across items per condition. Top right. 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. Results are similar as for the other semantic illusion simulation:* $t_{1(9)} = 1.69$, $p = .38$, $t_{2(7)} = 12.67$, $p < .0001$, *for the comparison between congruent condition and semantic illusion;* $t_{1(9)} = 13.31$, $p < .0001$, $t_{2(7)} = 6.76$, $p < .001$, *for the comparison between semantic illusion and incongruent condition, and* $t_{1(9)} = 12.18$, $p < .0001$, $t_{2(7)} = 7.36$, $p < .001$, *for the comparison between congruent and incongruent condition. Bottom. Activation of the unit "pine" in response to the Agent and Patient probe while the model processes a sentence from this semantic illusion simulation, literally "The pine has the man watered." (i.e., "The pine has watered the man." with Dutch word order). As for the other semantic illusion, the model represents the pine as the patient instead of the agent of the event throughout the sentence.*

Specificity of the N400 to violations of semantic rather than syntactic factors. While the N400 is sensitive to a wide range of semantic variables, amplitudes are not influenced by syntactic factors such as for instance violations of word order (e.g., "The girl is very satisfied with the ironed neatly linen.") which instead elicit P600 effects[36]. Because the model is representing the event described by the sentence, and this event itself is not necessarily affected by a change in word order, the model is likewise insensitive to such violations. To demonstrate this, we examined the model's response to changes in the usual word order (e.g., "On Sunday, the man *the robin* feeds" compared to "On Sunday, the man *feeds* the robin), examining the size of the semantic update at the highlighted position, where the standard word-order is violated. We found that, if anything, SU was actually slightly larger in the condition with the normal as compared to the changed word order (please see Fig. 4 and Supplementary Fig. 3; significant over models but not items). This is because changes in word order also entail changes in the amount of information a word provides about the event being described; it turns out that the amount of semantic update was on average slightly larger in the sentences with normal compared to changed word order (see *online methods* for details).

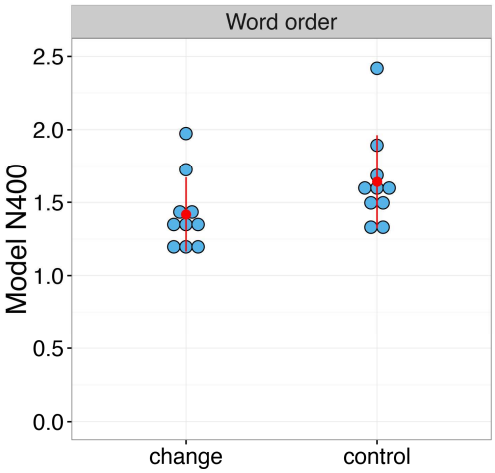
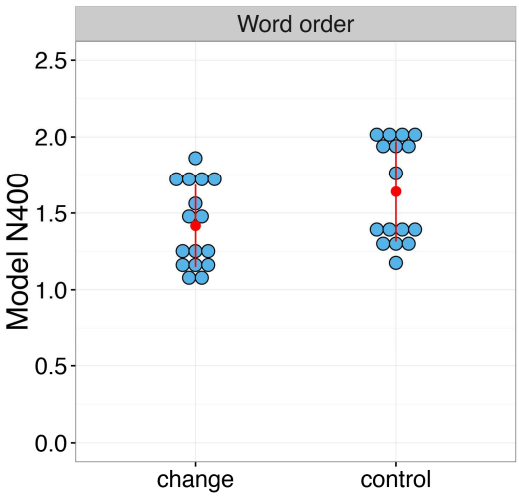


Figure 4. *Simulation of the influence of a change in normal word order. Change, changed word order; control, normal word order. Each blue dot represents the results for one independent run of the model, averaged across items per condition; the red dots represent the means for each condition, and red error bars represent +/- SEM. Semantic update was slightly larger for normal compared to changed word order; the main effect was significant over models,* $t_{1(9)} = 5.94$ *,* $p < .001$*, but not over items,* $t_{2(9)} = 1.56$*,* $p = .14$*.*



Supplementary Figure 3. *Simulation of the influence of a change in word order (by item). Change, changed word order; control, normal word order. Each blue dot represents the results for one item, averaged across 10 runs of the model; red dots represent means for each condition, and red error bars represent +/- SEM. Statistical results are reported in the caption of Fig. 4.*

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 that latently represents these estimates in a way that only becomes explicit when queried. In this section, we show that the implicit semantic update (measured at the hidden SG layer) and the change in the networks' explicit estimates of feature probabilities in response to probes (measured at the output layer) 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 a 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[37] but later decrease with age[38,39]. A comparison of the effect of semantic congruity on SU at different points in training shows a developmental pattern consistent with these findings (Fig. 5, top, and Supplementary Fig. 4, 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 output activation (Fig. 5, bottom, and Supplementary Fig. 4, bottom). The activation pattern at the output layer directly reflects explicit estimates of semantic feature probabilities in that units at the output layer explicitly specify semantic features, such as for instance “can grow”, “can fly” etc., and network error (across the training environment) is minimized when the activation of each feature unit in each situation corresponds to the conditional probability of this feature in this situation (e.g., an activation state of .7 in a situation where the conditional probability of the feature is .7). Thus, in the trained model, changes in output activation induced by an incoming word approximate changes in explicit estimates of semantic feature probabilities induced by that word. The continuing increase of the congruity effect across training displayed at the bottom of Fig. 5 thus shows that changes in explicit estimates of semantic feature probabilities do not pattern with the developmental trajectory of N400 effects. Instead, the change in hidden SG layer activation patterns with the N400 (Fig. 5, top), showing that the implicit and ‘hidden’ character of the model’s N400 correlate is crucial to account for the empirical data. The *decrease* in the amount of activation change at the hidden SG layer and the corresponding *increase* in the amount of activation change at the output layer over the later phase of learning shows that, as learning proceeds, less change in activation at the SG layer is needed to effectively support larger changes in explicit probability estimates. This pattern is possible because, as noted above, 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 become 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[40].

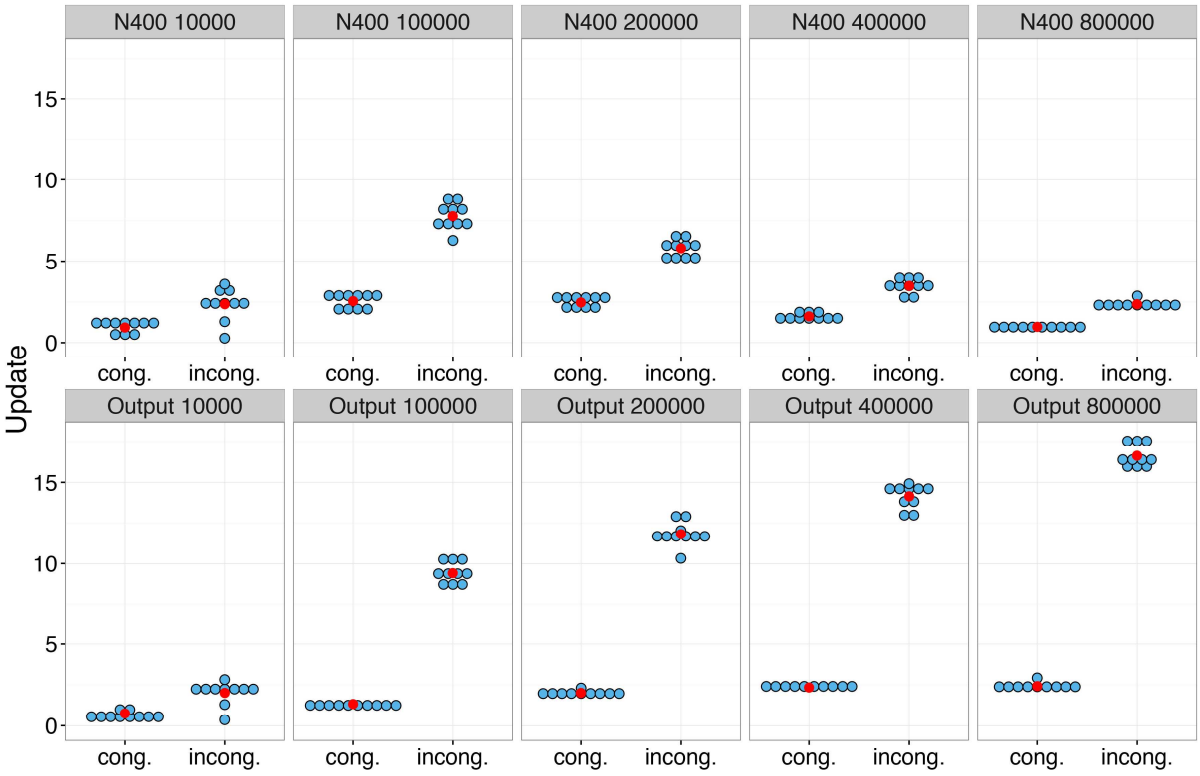
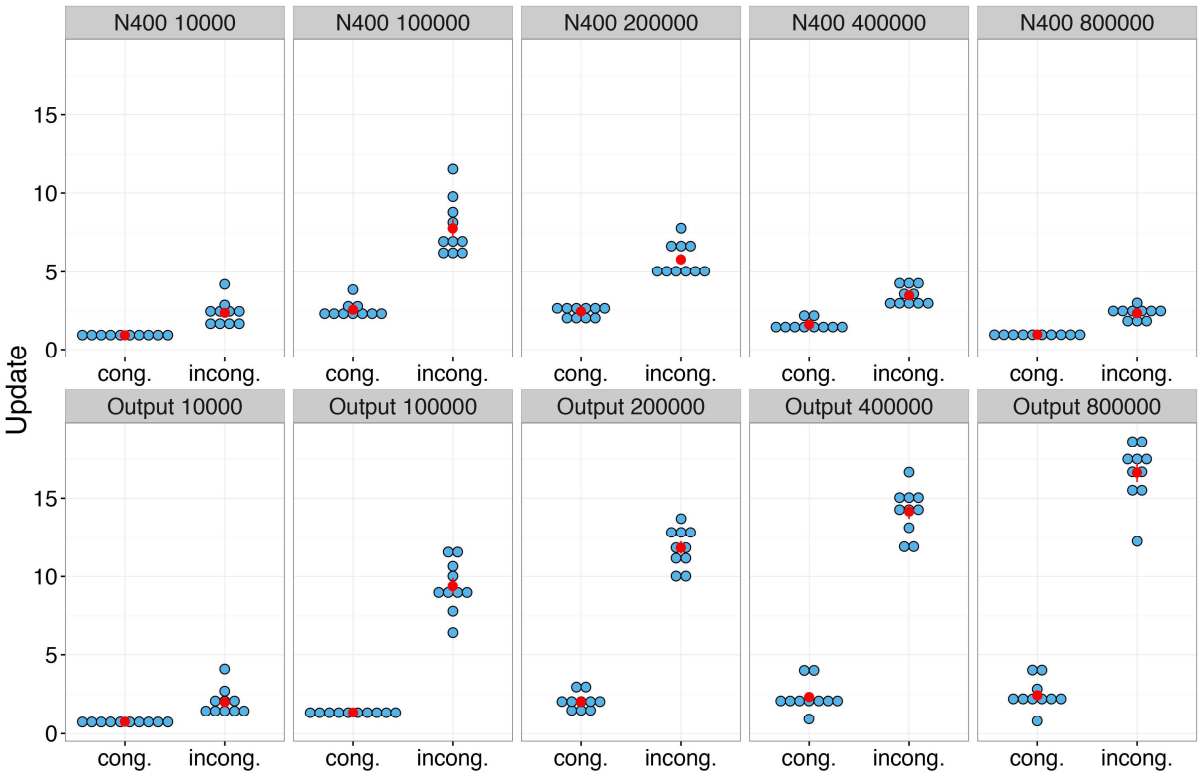


Figure 5. *Development across training. Semantic incongruity effects as a function of the number of sentences the model has been exposed to. Top. Semantic update at the model's hidden Sentence Gestalt layer shows at first an increase and later a decrease with additional training, in line with the developmental 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 4. *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. 5 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[41]. 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. 6a). Even at this early stage, SU was significantly influenced by semantic priming, associative priming, and semantic congruity in sentences (Fig. 6b and Supplementary Fig. 5) while overt estimates of feature probabilities were only weakly modulated by the words presented.

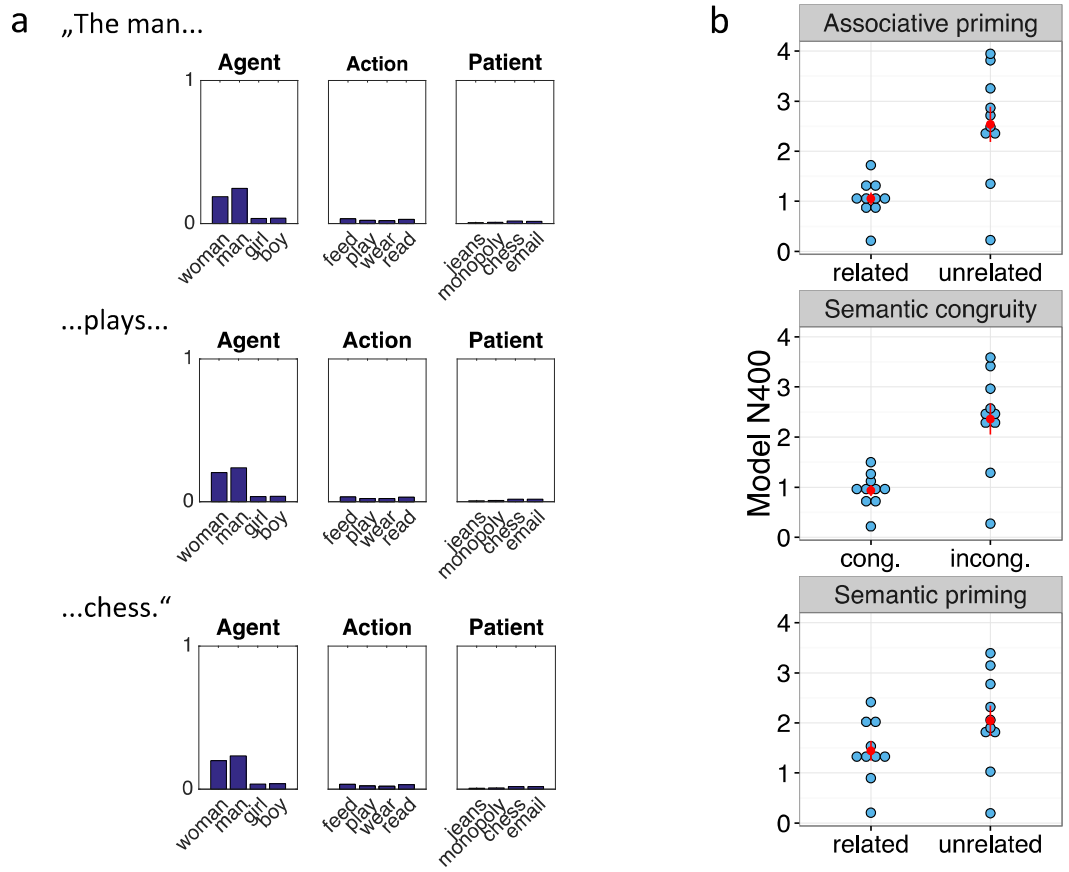
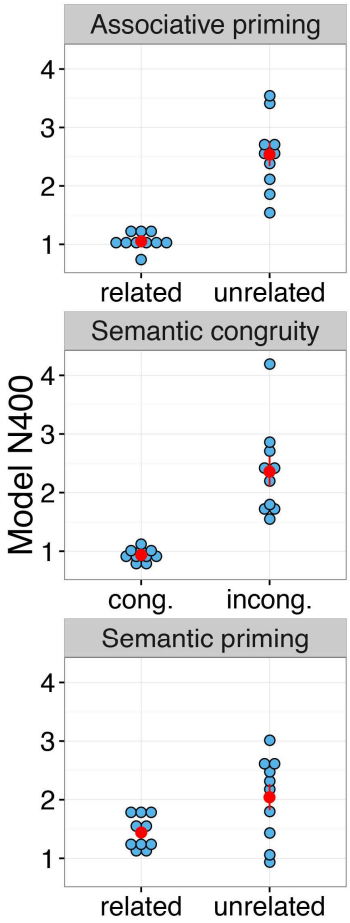


Figure 6. *Comprehension performance and semantic update effects at a very early stage in training. Cong., congruent; incong., incongruent. a. Activation of selected output units while the model is presented with the sentence "The man plays chess.". It can be seen that the model fails to activate the 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 5 (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. 6 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 estimate 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[42–44] 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: 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 that we back-propagate through the network to drive the adaptation of connection weights after each presented word becomes the difference in SG layer activation between the current and the next word, i.e. $SG_{n+1} - SG_n$ (see *online methods* for more details). 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[45].

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. 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. 7 and Supplementary 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[46].

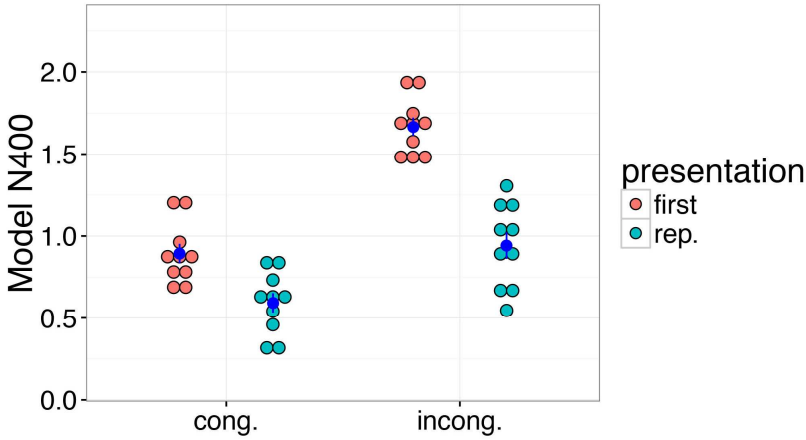
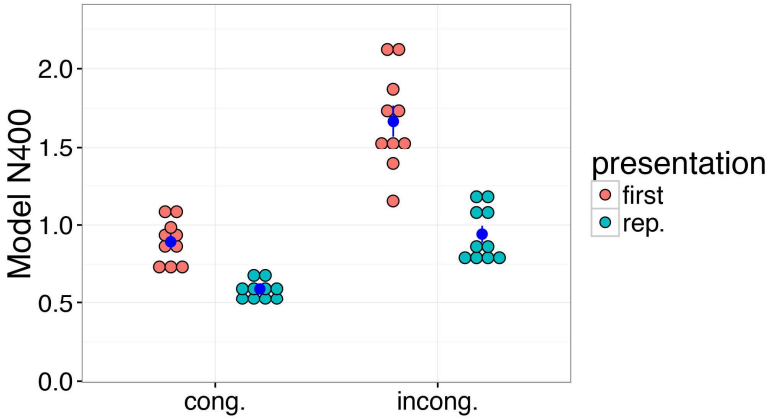


Figure 7. *Simulation of the interaction between delayed repetition and semantic incongruity. Cong., congruent; incong., incongruent; rep., repeated. Each red or green dot represents the results for one independent run of the model, averaged across items per condition; the blue dots represent the means 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 6. *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. 7.*

Discussion

The N400 ERP component is widely used to investigate the neurocognitive processes underlying the processing of meaning in language. As noted above, attempts to understand the factors that affect the N400 in terms of verbally formulated descriptive accounts grounded (at least in part) in traditional theories of language processing[8,10,11] have not fully succeeded in providing an adequate characterization of its functional basis. In the simulations presented above, we have shown that an implemented computational model that is grounded in an alternative approach to the nature of the language understanding process can provide a unified account that captures a wide range of findings (Table 1). The model treats N400 amplitudes as indexing the change induced by an incoming word in an implicit probabilistic representation of meaning in a neural network model that does not implement any of the existing descriptive accounts. The distinctive characteristics of the model that were described in the introduction are essential to its ability to account for the findings we have considered, as we explain below.

First, our model does not assume separate stages of lexical access/retrieval of word meanings and subsequent integration into a compositional representation. This is crucial because the two most prominent competing theories of the N400's functional basis suggest that N400 amplitudes reflect either lexical access[3] or integration (also referred to as unification) into a compositional (sometimes called combinatorial) representation of the meaning of the sentence [6,7]. In the SG model, incoming stimuli instead serve as 'cues to meaning'[47] which automatically change an activation pattern that implicitly represents conditional probabilities of all aspects of meaning. Our account is similar to the lexical access perspective in that the process is assumed to be fast, automatic, and implicit, but differs from this view in that the resulting activation pattern represents not just the currently incoming word but instead corresponds to an updated implicit representation of the event being described by the sentence. In this regard our account is similar to the integration view in that

the resulting activation state is assumed to represent all aspects of meaning of the described event (including – though this aspect is not currently implemented – input from other modalities), though it differs from such accounts in avoiding a commitment to explicit compositional representation. Our perspective seems in line with a recent comprehensive review on the N400 ERP component[2] which concluded that the N400 "does not readily map onto specific subprocesses posited in traditional frameworks" (p. 639) 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. 641). Crucially, the SG model provides a computationally explicit account of the nature and role of this distributed activation state and how it changes through stimulus-driven activity as meaning is dynamically constructed during comprehension. Because the model uses incoming words as cues to semantic event features instead of linguistic representations in which words are placed into specific syntactic roles, it does not predict an N400 response to reversal anomalies (Fig. 2i & Supplementary Fig. S2) or to violations of word order (Fig. 4).

Second, the model does not specify a specific structure of the model's internal representations. Instead the representations result from a learning process and thus depend on the statistical regularities in the model's environment as well the amount of training the model has received, allowing it to account for the pattern of N400 effects across development (Fig. 5) including N400 effects while behavioral performance is still near chance (Fig. 6) as well as the influence of relatively long-term repetition on N400 congruity effects (Fig. 7).

Third, the model updates its activation pattern upon the presentation of any word, allowing it to capture N400 effects for single words (i.e., frequency effects; see Fig. 2e) and words presented in pairs (influences of repetition, Fig. 2h, semantic priming, Fig. 2f, and associative priming, Fig. 2g) as well as words presented in a sentence context (influences of semantic congruity, Fig. 2a, cloze probability, Fig. 2b, position in the sentence, Fig. 2c, and semantically related incongruity, Fig. 2d).

Fourth, the N400 as captured by the model is assumed to characterize one specific aspect of language comprehension, namely the automatic stimulus-driven update of an initial implicit representation of meaning. This characterization is in line with evidence for the N400's anatomical localization in regions involved in semantic representation such as the medial temporal gyrus (MTG[3]) and anterior medial temporal lobe (AMTL[48,49]). The processes underlying the N400 may thus correspond to the type of language processing that has been characterized as sometimes shallow[50] and "good enough"[54] and that is preserved in patients with lesions to frontal cortex (specifically left inferior prefrontal cortex, BA47)[52,53]. Thus, activity in temporal lobe regions MTG and AMTL may correspond to immediate, automatic, and implicit aspects of sentence processing as captured by the SG model. In contrast, the left, inferior frontal cortex has been proposed to support control processes in comprehension that are required only when processing demands are high[54,55] such as in syntactically complex sentences[52] which require selection among competing alternatives[56]. These aspects of language comprehension may be reflected in other ERP components as discussed below.

The pattern of activation in the model's Sentence Gestalt (SG) layer latently predicts the 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. While in the current implementation of the model, inputs are presented over a series of discrete time steps corresponding to each successive word in the sentence, this is just a simplification for tractability. We assume that in reality, the adjustment

of the semantic activation occurs continuously in time as auditory or visual language input is processed, so that the earliest arriving information about a word (whether auditory or visual) immediately influences the evolving SG representation[57]. This assumption is in line with the finding that N400 effects in spoken language comprehension often begin to emerge before the spoken word has become acoustically unique[58,59]. It is important to underline the point that this kind of prediction does not refer to the active and intentional prediction of specific items but rather to a latent or implicit state such that the model (and presumably the brain) becomes tuned through experience to anticipate likely upcoming input to respond to it with little additional change. 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 the state of the world conveyed by the sentence and at the same time correspond to the learning signal driving adaptation of connection-based knowledge representations. In this sense, our approach, first introduced almost 30 years ago, anticipates predictive coding approaches to understanding the dynamics of neural activity patterns in the brain [43,60] . Our simulations suggest that the semantic system may not represent probabilities of aspects of meaning explicitly but rather uses a summary representation that implicitly represents estimates of these probabilities, supporting explicit estimates when queried and 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[14–16,61]. 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[26]. Because this 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. To demonstrate this, we trained an SRN on the same training corpus as the SG model and repeated some of the critical simulations with this SRN (Fig. 8 and Supplementary Fig. 7; see *online methods* for details).

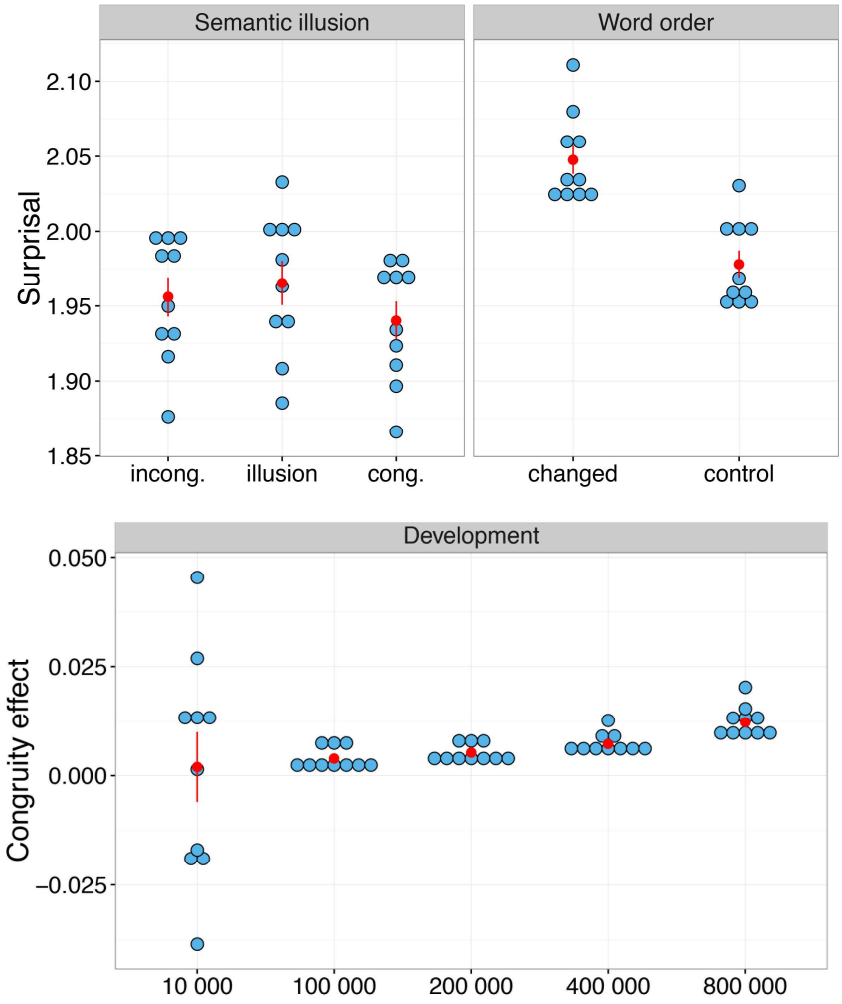
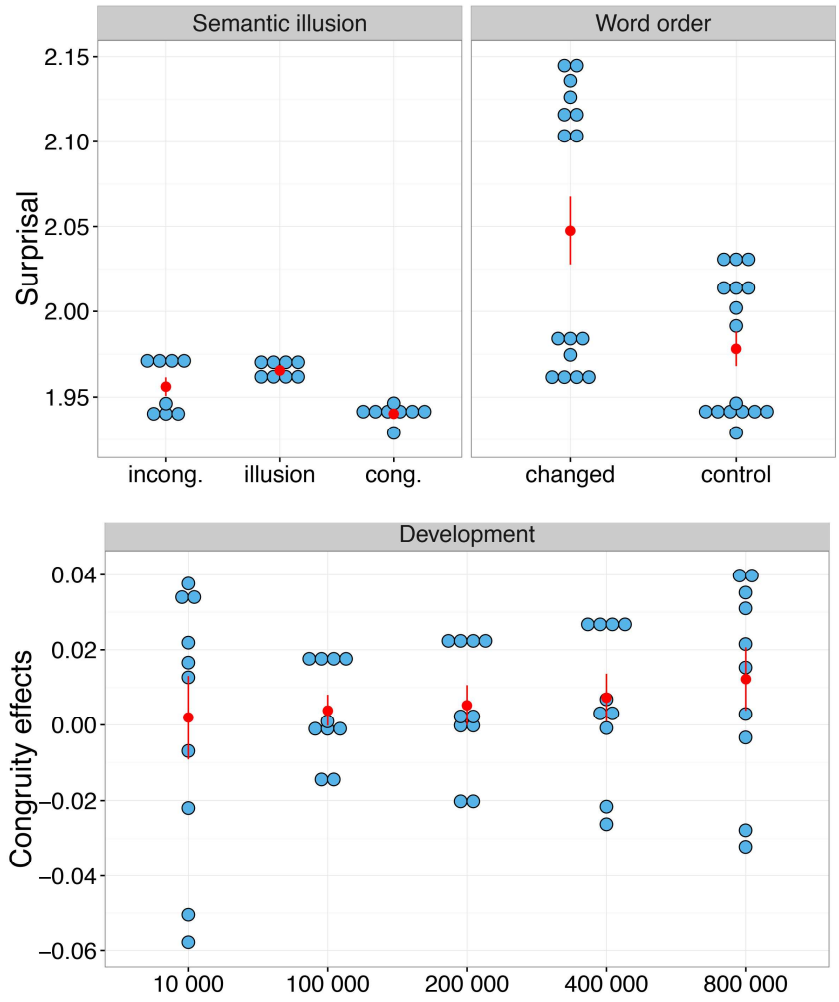


Figure 8. *Simulation results from a simple recurrent network model (SRN) trained to predict the next word based on the preceding context. Each blue dot represents the results for one independent run of the model, averaged across items per condition; the red dots represent the means for each condition, and red error bars represent +/- SEM. Top left, semantic illusion:* $t_{1(9)} = 4.55$, $p < .01$, $t_{2(7)} = 7.83$, $p < .001$ *for the comparison between congruent and illusion condition;* $t_{1(9)} = 12.28$, $p < .0001$, $t_{2(7)} = 2.98$, $p = .062$ *for the comparison between congruent and incongruent condition;* $t_{1(9)} = 1.52$, $p = .49$, $t_{2(9)} = 1.57$, $p = .48$ *for the comparison between incongruent and illusion condition. Top right, word order:* $t_{1(9)} = 29.78$, $p < .0001$; $t_{2(15)} = 6.73$, $p < .0001$. *Bottom, congruity effect on surprisal as a function of the number of sentences the model has been exposed to:* $t_{1(9)} = .26$, $p = 1.0$, $t_{2(9)} = .15$, $p = 1.0$ *for the comparison between 10 000 and 100 000 sentences;* $t_{1(9)} = 6.74$, $p < .001$, $t_{2(9)} = 1.08$, $p = 1.0$ *for the comparison between 100 000 and 200 000 sentences;* $t_{1(9)} = 7.45$, $p < .001$, $t_{2(9)} = 1.78$, $p = .44$ *for the comparison between 200 000 and 400 000 sentences;* $t_{1(9)} = 10.73$, $p < .0001$, $t_{2(9)} = 1.93$, $p = .36$ *for the comparison between 400 000 and 800 000 sentences.*



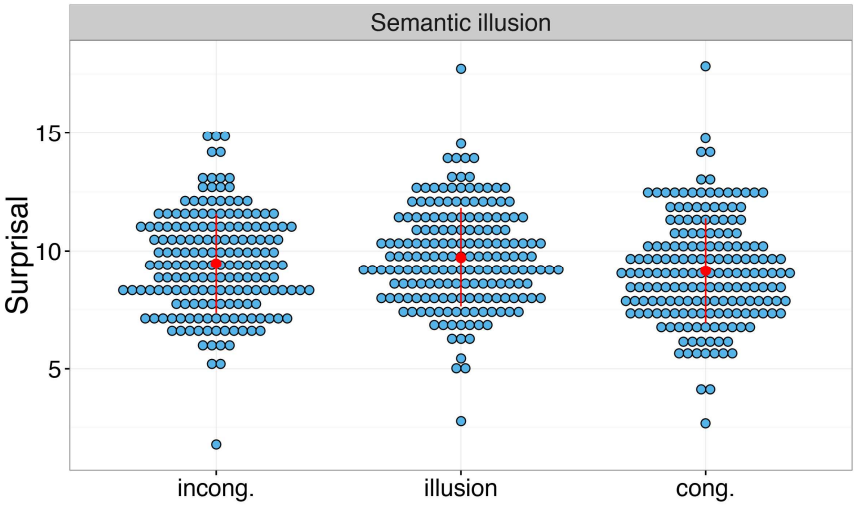
Supplementary Figure 7. *Simulation results from a simple recurrent network model (SRN) trained to predict the next word based on the preceding context. Each blue dot represents the results for one item, averaged across 10 runs of the model; red dots represent means for each condition, and red error bars represent +/- SEM. Statistical results are reported in the caption of Fig. 8. Top left, semantic illusion. Top right, word order. Bottom, congruity effect on surprisal as a function of the number of sentences the model has been exposed to.*

First, word surprisal reflects both semantic and syntactic expectation violations, while the N400 is specific to semantic expectations as described above. Indeed, while SU in the SG model was insensitive to changes in word order (Fig. 4 and Supplementary Fig. 3), surprisal in the SRN was significantly larger for changed as compared to normal word order (see Fig. 8 and Supplementary Fig. 7). The lack of specificity of the word surprisal measure converges with the finding that the correlation between surprisal in the SRN and N400 observed in the above mentioned study [26] was observed only for content words; the SRN surprisal measure

when calculated over grammatical function words did not correlate with the N400 responses observed on these words.

Furthermore, the SRN did not account for the decrease of N400 effects with age, showing instead a slight increase with additional training (see Fig. 8 and Supplementary Fig. 7). This is because surprisal is measured in terms of the estimates of word probabilities, which become sharper as learning progresses. Finally, the SRN did not produce the small N400 in reversal anomalies: When presented with "At breakfast, the eggs *eat*…", word surprisal was large, numerically even larger than an incongruent continuation (see Fig. 8 and Supplementary Fig. 7) while semantic update in the SG model shows only a very slight increase, in line with N400 data[32] (see also Supplementary Fig. 8 and the accompanying text for relevant results from an SRN trained on a natural corpus by S. Frank (personal communication)).

The other sentence-level model focuses specifically on reversal anomalies, assuming separate stages of lexical retrieval and semantic integration [33]. This retrieval-integration model is computationally explicit while following aspects of the classical framework for language processing, in which there is thought to be a distinct lexical-semantic processing module in which spreading activation can occur among related items, prior to integrating the retrieved word meanings into a compositional representation of sentence meaning[9,63]. The retrieval-integration model makes the further assumption that reversal anomalies such as 'for breakfast the eggs would *eat'* must produce a large update in the representation of sentence meaning, since the sentence appears to describe an event in which eggs are agents engaged in the act of eating. 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.



Supplementary Figure 8. *Simulation results from a simple recurrent network (SRN) implementation by T. Mikolov*[62] *trained by S. Frank on 23M sentences from a web corpus. Incong., incongruent; cong., congruent. The simulation experiment consisted in the presentation of materials from the semantic illusion experiment by Kuperberg and colleagues*[32] *which we requested from the authors (there are a few slight differences in the materials due to an issue with retrieving the original stimuli, but the materials largely overlap and resulted in the same pattern of results; G. Kuperberg, personal communication). Each blue dot represents the results for one item, averaged over three runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM. Results resemble those from the SRN that we trained on the same corpus as the SG model (Fig. 8 and Supplementary Fig. S7) in that word surprisal was large in the semantic illusion condition, numerically even larger than in the incongruent condition. There were 3 runs of the model and 180 items in each condition (1 less in the incongruent condition because the model did not know one of the words in this condition, "curtseys") so that we report statistical results from the item analyses:* $t_{2(179)} = 11.76$, $p < .0001$, *for the comparison between congruent condition and semantic illusion;* $t_{2(178)} = 1.29$, $p = .59$, *for the comparison between semantic illusion and incongruent condition, and* $t_{2(178)} = 1.45$, $p = .45$, *for the comparison between congruent and incongruent condition. We thank Stefan Frank for performing the simulation and sharing the results with us!*

As discussed above, 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 wide range of N400 effects which traditional accounts would ascribe either to lexical access or to subsequent semantic integration. Crucially, our model accounts for the absence of an N400 in reversal anomalies because such sentences do not trigger a re-assignment of the role of eggs in a compositional representation of the meaning of the sentence; instead the implicit internal representation continues to treat the eggs as having been

eaten, consistent with the model's knowledge of the constraints affecting natural events, in which eggs lack attributes that would allow them to serve as agents of eating.

While both the retrieval-integration model and the SG model account for the absence of N400's in reversal anomalies, the SG model does so within the context of a more complete account of the factors that do and do not influence the N400, while the retrieval-integration model has yet to be extended beyond accounting for a subset of the relevant findings. Further research will be required to determine whether the retrieval-integration model can be extended to encompass the range of N400 findings encompassed by the SG model. There are also challenges to the view that the P600 should be thought of as reflecting the process of semantic integration as it ordinarily occurs in language processing, as we discuss below.

One basic challenge to the retrieval-integration model's claim that the P600 reflects semantic integration is the fact that many variables that should influence the amount of change in a representation of sentence meaning, such as cloze probability or surprise, consistently influence N400 but do not necessarily influence P600 amplitudes[23,26,64]. Some studies report influences of cloze probability on a post N400 positivity[64–66], and this finding might be taken as consistent with the retrieval-integration model. However, these influences consistently show a frontal topographical distribution, different from the parietal P600 effects obtained in reversal anomalies and related materials, leading many researchers to suggest that these two ERP positivities reflect functionally distinct processes[64]. Furthermore, the influence of cloze probability on the frontal positivity does not seem to be influenced by degree of semantic similarity, as would be expected if it was related to the change in a representation of sentence meaning. Instead, the effect is dichotomous, that is, larger for unexpected words independent of their semantic similarity with expected words[65] which has been taken to suggest that the frontal positivity reflects specific lexical predictions. These findings appear to challenge any model linking the P600 (without further differentiation) to a change in the representation of sentence meaning [33].

The functional basis of the late positive ERP components we have described is not addressed by our model and requires further investigation to be more fully understood. It is true that P600 responses have been observed to a wide range of linguistic violations and irregularities, including reversal anomalies[32,35], syntactic violations[36], and garden path sentences[67], as well as pragmatic processes (see review[68]). This has been taken to suggest that the P600 might reflect combinatorial aspects of language processing, either related to syntax[36] or to semantic integration as assumed in the retrieval-integration model[33]. There is, however, an alternative perspective, in which the P600 is not treated as specific to language processing (either syntactic processes or semantic integration) *per se*, but to a more general process that may be associated with more conscious, deliberate, and effortful aspects of processing which may result in adjustments to the initial implicit representation of meaning reflected in the N400. Several researchers have pointed out that the P600 shares properties with the P3[69,70] which is elicited by the occurrence of oddball stimuli (such as a rare high tone among much more frequent low tones), with the component's latency depending on stimulus complexity. This component is thought to signal an explicit surprise response and a corresponding update in working memory[71]. This P600-as-P3 perspective naturally explains the observed sensitivity of P600 effects to task demands and attentional focus. Indeed, P600 effects are strongly reduced or absent when there is no active task or when the task is unrelated to the linguistic violation[72]. In contrast, N400 effects can be obtained during passive reading and even during unconscious processing such as within the attentional blink[73]. Thus, from this view, the P600 differs from the N400 in two ways. It belongs to a component family that responds to a wider range of expectation violations while the N400 is specific to the formation of a representation of meaning. Further, the N400 may reflect an automatic and implicit process while the P600 may be associated with a higher level of control and attention, allowing it to be affected by additional constraints that the semantic update process underlying N400 amplitudes misses

out on. As noted above, these issues should be investigated in future research to be more fully understood.

In general, the current work opens up an opportunity for extensive further investigations, addressing a wide range of behavioral as well as neural aspects of language processing. One interesting finding that should be addressed in future work is the finding that N400s were influenced by categorical relationship (i.e., semantic priming, see Fig. 2f) while being unaffected by sentence truth, at least in negated statements: The N400 is equally small in the false sentence "A robin is not a bird" and the true sentence "A robin is a bird", and is equally large in the true sentence "A robin is not a vehicle" and the false sentence "A robin is a vehicle" [74]. It is important to note that sentence truth is not the same as expected sentence meaning, and that to understand the influence of negation on meaning expectations, one needs to take into account the pragmatics of negation[75,76]. Specifically, negation is typically used to deny a supposition, and in the absence of discourse context, this supposition must be grounded in general knowledge[75]. Thus, when used in short and isolated sentences, negation is typically used to deny something that is part of an invoked schema (e.g., "a whale is not a fish"). "Robin" does not invoke a schema which includes semantic features of "vehicle" so that "A robin is not a vehicle" is not an expected sentence meaning, even though it is true. On the other hand, "robin" does invoke a schema which includes semantic features of "bird" so that something that is part of the schema of "bird" might be expected to be denied (e.g., "A robin is not a bird that flies south during winter" is fine). Follow-ups taking the pragmatics of negation into account and providing more context showed that N400s are indeed modulated by sentence truth[76] and plausibility[75]. Our model currently has no experience with sentences that describe properties of classes of objects, as sentences like 'a whale is not a fish' do, but such sentences could be incorporated in an extension of the model, allowing further research to investigate whether the pattern of semantic update seen in such sentences can be captured by our account of N400 amplitudes as change in a probabilistic representation of meaning.

Furthermore, it remains to be explored how well the SG model can predict behavioral measures of sentence processing. Given the extensive evidence reviewed above that the update of an implicit probabilistic representation of meaning is only one of the processes that occur during language processing, it seems likely that a full account of overt behavioral responses would require a fuller model capturing these other processes. The beauty of ERPs is that they may index distinct aspects of these processes, and can thus speak to their neurocognitive reality even though several such processes might jointly influence a specific behavioral measure. To fully address behavior, the model will likely need to be integrated into a more complete account of the neuro-mechanistic processes that take place during language processing, including the more controlled and attention-related processes that may underlie the P600. In addition, the model's query language and training corpus will need to be extended to address the full range of relevant phenomena, including other ERP components (e.g., orthographic and syntactic ERPs) as well as signals that have been detected using other measurement modalities[55,77].

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, departing from constructs postulated by traditional language processing theories[8,10,11] that have lingered on in many previous accounts of the functional basis of the N400. The 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 (SG) 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. Finally, we describe simulations conducted with an SRN. Figure 1 in the main text presents the SG network architecture and the processing flow in the model.

Implicit probabilistic theory of 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 - in the general form of the theory - specific bounds set on the content of the representation of meaning. In any specific implementation of the theory, the content of the representation of meaning is prescribed by the range of possible probes and queries, which in the case of our implementation correspond to the vectors encoding the pairs of thematic roles and their fillers. 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. The query-answer form is used instead of directly providing the complete event description at the output layer to keep the set of probes and fillers more open-ended and to suggest the broader framework that the task of sentence comprehension

consists in building internal representations that can be used as a basis to respond to probes[12]. 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[13,78,79], 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[13]. 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[13].

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[80], 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[17,25]. In the framework of our approach to the characterization of sentence

meaning, we adapt one of these proposals[17], 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:

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

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:

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

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*[81] 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[82] 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. The model would not be able to learn language based on this semantic update driven learning rule alone. We assume that language learning proceeds by a mixture of experience with language processed in the context of observed events (as in the standard training regime) and processed in isolation (as with the semantic update driven learning rule), possibly with changing proportions across development. Future modeling work should explore this issue in more detail.

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. 9a). The sentences are single clause sentences such as "At breakfast, the man eats eggs in the kitchen".

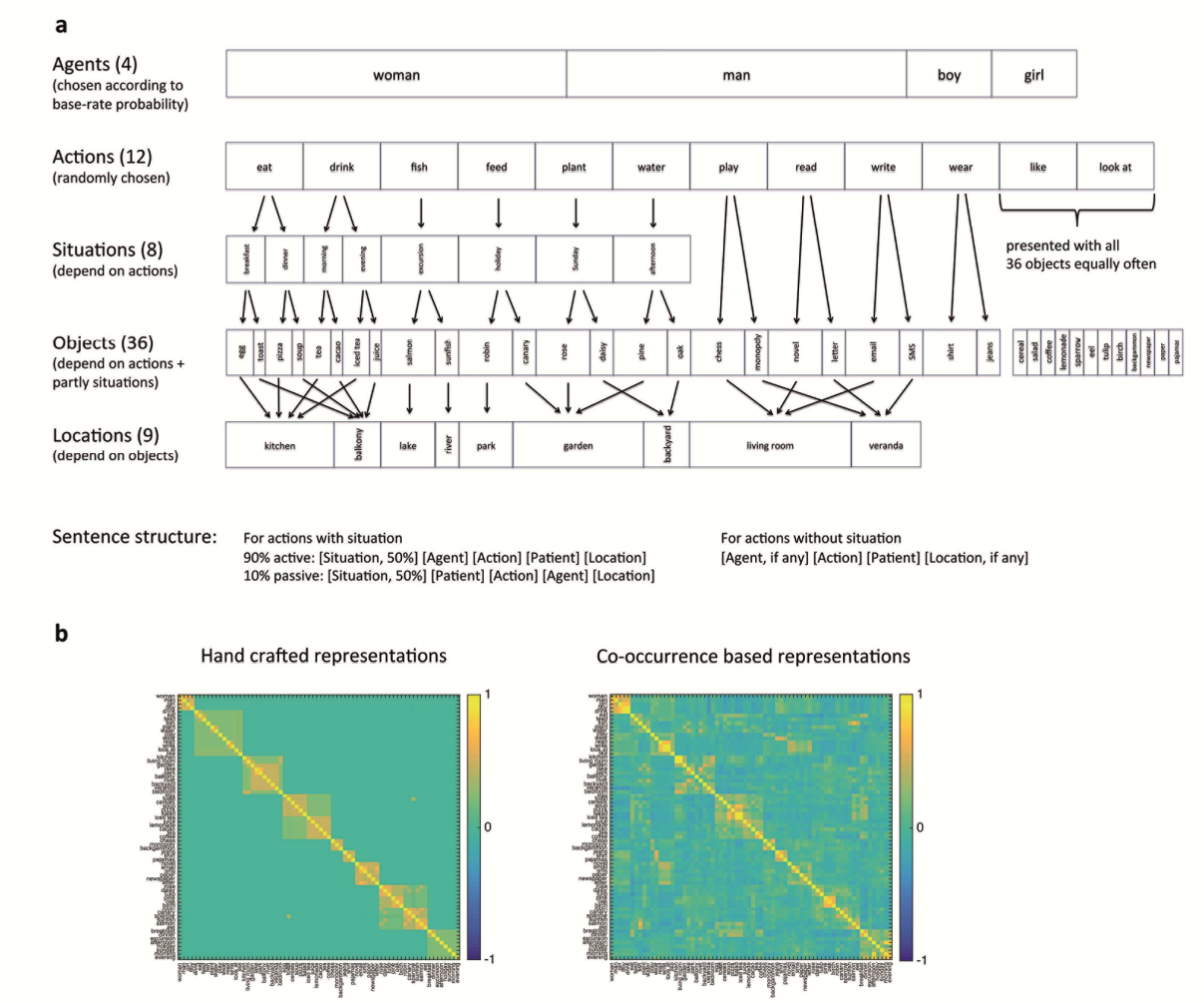


Figure 9. *a. The sentence/ event generator used to train the model. Bar width corresponds to relative probability. First, one out of twelve actions is chosen with equal probability. Then, for every action except one ("look at") an agent is chosen ("woman" and "man" each with a probability of .4, "boy" 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*[83]*. The correlation between the matrices is r = .73.*

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".

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[84,85]. 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[83]

showed a reasonable correspondence (r = .73; see Fig. 9b), 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 approximates 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. It is important to note that this is not meant to be a principled theoretical assumption but is rather just a practical consequence of the training approach. In general, we do not assume that listeners can only learn when they simultaneously experience a described event, first, because neural networks can generalize[12] and second, because the SG model can also learn simply from listening or reading based on the new learning rule driven by the semantic update (see section *Semantic update driven learning rule*, above). Also, observed events can be ambiguous and language can provide a particular disambiguating perspective on an event that cannot be gleaned directly from the event itself[86]. The SG model implements a simplification of the situation in the sense that events in the model are always unambiguous and complete. In addition, the training procedure implements the assumption that listeners anticipate the full meaning of each presented sentence as early as possible[87,88], 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. 9a, 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[89]. 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[32] 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.

We also simulated a second type of semantic illusion where a relationship between two noun phrases is established prior to encountering the verb[35] (e.g. "De speer heft de atleten *geworpen*", lit: "The javelin has the athletes *thrown*", relative to "De speer werd door de atleten *geworpen*", lit: "The javelin was by the athletes *thrown*"). For this simulation we presented basically the same stimuli as for the other semantic illusion simulation, but with Dutch word order and thus sentence structures relevant to examine whether the same mechanism allowing the model to account for the semantic illusion effects reported by Kuperberg et al. would also hold when the verb is presented at the end of the sentence. Thus, the relevant experimental conditions contained sentences such as "The pine was by the man watered." (i.e., "The pine was watered by the man." with Dutch word order; congruent condition), "The pine has the man watered." (i.e. "The pine has watered the man." with Dutch word order; semantic illusion/ reversal anomaly condition) and "The pine was by the man drunken." (i.e., "The pine was drunken by the man." with Dutch word order; incongruent condition). To be able to run this simulation, we trained a model on basically the same training environment as the other model, but with the sentence structures adjusted such that active sentences were changed from e.g., "The man waters the pine." to "The man has the pine watered." and passive sentences were changed from "The pine was watered by the man." to "The pine was by the man watered.". We also added an additional input unit representing "has" and made "was by" be represented by a single unit because both words now always occurred in direct succession (e.g., "… *was by* the man watered." instead of "… *was* watered *by* the man."). Apart from that, all parameters of the model and training were kept the same. This implementation does not completely correspond to the empirical experiment[35] in that in

our simulation there was no specific relationship between the agent and the action (i.e., the man in the model environment is equally likely to perform all 12 actions and thus was equally likely to water something as he was to drink something, for instance) while in the stimulus material of the empirical experiment there was a specific probabilistic relationship between the agents and the actions (i.e., athletes might be more likely to throw something than to summarize something). However, important for current purposes, this implementation allowed to test whether the way the model accounts for the slight N400 increase in reversal anomalies would be robust to changes in word order, i.e. the presentation of two noun phrases prior to the presentation of the verb. For the simulation, there were eight items in each experimental condition, and semantic update was computed as the difference in SG layer activation between the third constituent ("man", word *n*-1) and the fourth constituent (the action, word *n*).

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[42–44]. 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*).

For the simulation of the influence of violations of word order (phrase structure)[36], we presented two types of word order changes for each sentence, focusing on sentences starting with a situation, because in these sentences it is easier to keep changes in conditional probabilities of semantic event features relatively low when changing word order. For each sentence, we presented (1) a version where we changed the position of the action and the patient (e.g., "On Sunday, the man *the robin* feeds" compared to "On Sunday, the man *feeds* the robin"; with semantic update computed as the difference in SG layer activation between the presentation of the agent (word *n*-1) and the patient or action, respectively (word *n*)), and (2) a version where we changed the position of the agent and the action (e.g., "On Sunday, *feeds* the man the robin" compared to "On Sunday, *the man* feeds the robin"; with semantic update computed as the difference in SG layer activation between the presentation of the situation (word *n*-1) and the action or agent, respectively (word *n*)). For type (1), changing position of action and patient, the conditional probability of the semantic features associated with the critical word (not at this position in the sentence but in general within the described event) is .7 in the condition with the changed word order and 1.0 in the condition with the normal word order. For type (2), changing position of agent and action, the conditional probability of the semantic features associated with the critical word (again, crucially, not at this position in the sentence but in general within the described event) is 1.0 in the condition with the changed word order and .4 in the condition with the normal word order. Thus, while changes in word order also entail changes in the amount of semantic update of event features, the design of the simulation ensures that influences of word order (syntax) and semantic

update can be dissociated. Specifically, the surprise concerning the semantic features of the described event was on average .15 in the condition with the changed word order (.3 for type (1) and 0.0 for type (2)) while it was on average .3 in the condition with the normal word order (0.0 for type (1), and .6 for type (2)). There were 16 items (8 of each type) in each condition (i.e., normal vs. changed word order).

Simple recurrent network model simulations

We trained a classic simple recurrent network[90] (consisting of an input and output layer with 74 units each, as well as a hidden and context layer with 100 units each) on the same training corpus as the SG layer. Except for the architectural difference, all parameters were kept the same. We then simulated influences of violations of word order (phrase structure), reversal anomalies, and development, as described above for the SG model. The measure for surprisal that we set in relation to N400 amplitudes consists in the summed magnitude of the cross-entropy error induced by the current word (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 16; please see the previous section for the exact number of items in each simulation experiment) treated as random factor. There is much less noise in the simulations as compared to empirical experiment such that the relatively small sample size (10 runs of the model and 8 to 16 items per condition) should be sufficient. 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 > .24 for the item analyses) except for the item analysis of the change in word order ($p = .048$) which might be due to the items in this simulation experiment consisting of two types with slightly different characteristics (see section *Simulation of empirical findings* above); this item analysis did not reach significance neither in the t-test (see caption of Fig. 4) nor in the Wilcoxon signed rank test ($p = .10$) which does not depend on the normality assumption. 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; 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 in the SG model (Supplementary Fig. 1i; $p = .012$) as well as in the SRN (Supplementary Fig. 7; $p = .043$), and for the difference between changed and normal word order in the SRN (Supplementary Fig. 7;

p = .011). Again, Wilcoxon signed rank tests confirmed significant differences between the incongruent completions and the semantic illusions in the SG model (Supplementary Fig. 1i; p = .0078) and the SRN (Supplementary Fig. 7; p = .039), as well as a significant influence of word order in the SRN (p = .0004).

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. 1-8).

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. Debruille, J. B. The N400 potential could index a semantic inhibition. *Brain Res. Rev.* **56,** 472–477 (2007).

5. 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).

6. Baggio, G. & Hagoort, P. The balance between memory and unification in semantics: A dynamic account of the N400. *Lang. Cogn. Process.* **26,** 1338–1367 (2011).

7. Brown, C. & Hagoort, P. The processing nature of the N400: Evidence from masked priming. *J. Cogn. Neurosci.* **5,** 34–44 (1993).

8. Chomsky, N. *Syntactic structures*. (Mouton, 1957).

9. Fodor, J. *Modularity of Mind*. (MIT Press, 1981).

10. Fodor, J. & Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis. *Cognition* **28,** 3–71 (1988).

11. Jackendoff, R. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. (Oxford University Press, 2002).

12. McClelland, J. L., St. John, M. & Taraban, R. Sentence comprehension: A parallel distributed processing approach. *Lang. Cogn. Process.* **4,** 287–336 (1989).

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

14. 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).

15. 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).

16. Cheyette, S. J. & Plaut, D. C. Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition* **162,** 153–166 (2017).

17. Itti, L. & Baldi, P. Bayesian Surprise Attracts Human Attention. 1–8 (2006). doi:10.1016/j.visres.2008.09.007

18. Griffiths, T. L., Steyvers, M. & Tenenbaum, J. B. Topics in Semantic Representation. **114,** 211–244 (2007).

19. Andrews, M., Vigliocco, G. & Vinson, D. Integrating experiential and distributional data to learn semantic representations. *Psychol. Rev.* **116,** 463–498 (2009).

20. Wu, Y. *et al.* Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144* (2016).

21. Seidenberg, M. S. & McClelland, J. L. A Distributed, Developmental Model of Word Recognition and Naming. *Psychol. Rev.* **96,** 523–568 (1989).

22. McClelland, J. L. in *The Handbook of Language Emergence* (eds. MacWhinney, B. & O'Grady, W.) 54–80 (John Wiley & Sons, 2015).

23. Kutas, M. & Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature* **307,** 101–103 (1984).

24. Van Petten, C. & Kutas, M. Influences of semantic and syntactic context on open- and closed-class words. *Mem. Cogn.* **19,** 95–112 (1991).

25. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106,** 1126–1177 (2008).

26. 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).

27. 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).

28. Hagoort, P., Baggio, G. & Willems, R. M. in *The Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 819–836 (MIT Press, 2009).

29. Barber, H., Vergara, M. & Carreiras, M. Syllable-frequency effects in visual word recognition: evidence from ERPs. *Neuroreport* **15,** 545–548 (2004).

30. Koivisto, M. & Revonsuo, A. Cognitive representations underlying the N400 priming effect. *Cogn. Brain Res.* **12,** 487–490 (2001).

31. Rugg, M. D. The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology* **22,** 642–647 (1985).

32. 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).

33. 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.*

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

35. Hoeks, J. C. J., Stowe, L. A. & Doedens, G. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cogn. Brain Res.* **19,** 59–73 (2004).

36. Hagoort, P. & Brown, C. M. ERP effects of listening to speech compared to reading: the P600 / SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia* **38,** 1531–1549 (2000).

37. 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).

38. 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).

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

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

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

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

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

44. 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).

45. 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).

46. 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).

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

48. McCarthy, G., Nobre, A. C., Bentin, S. & Spencer, D. D. Language-Related Field Potentials in the Anterior-Medial Temporal Lobe: I. Intracranial Distribution and Neural Generators. *J. Neurosci.* **15,** 1080–1089 (1995).

49. Nobre, A. C. & McCarthy, G. Language-Related Field Potentials in the Anterior-Medial Temporal Lobe: II. Effects of Word Type and Semantic Priming. *J. Neurosci.* **15,** 1090–1098 (1995).

50. Sanford, A. J. & Sturt, P. Depth of processing in language comprehension: Not noticing the evidence. *Trends Cogn. Sci.* **6,** 382–386 (2002).

51. Ferreira, F., Bailey, K. G. D. & Ferraro, V. Good-Enough Representations in Language Comprehension. *Curr. Dir. Psychol. Sci.* **11,** 11–15 (2002).

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

53. Turken, A. U. & Dronkers, N. F. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Front. Syst. Neurosci.* **5,** 1–20 (2011).

54. Bookheimer, S. Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annu. Rev. Neurosci.* **25,** 151–188 (2002).

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

56. Thompson-Schill, S. L., D'Esposito, M., Aguirre, G. K. & Farah, M. J. Role of left inferior prefrontal cortex in retrieval of semantic knowledge : A reevaluation. *Proc. Natl. Acad. Sci.* **94,** 14792–14797 (1997).

57. Clayards, M., Tanenhaus, M. K., Aslin, R. N. & Jacobs, R. A. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* **108,** 804–809 (2008).

58. Van Petten, C., Coulson, S., Rubin, S., Plante, E. & Parks, M. Time course of word identification and semantic integration in spoken language. *J. Exp. Psychol. Learn. Mem. Cogn.* **25,** 394–417 (1999).

59. van den Brink, D., Brown, C. M. & Hagoort, P. The Cascaded Nature of Lexical Selection and Integration in Auditory Sentence Processing. *J. Exp. Psychol. Learn. Mem. Cogn.* **32,** 364–372 (2006).

60. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature* **2,** 79–87 (1999).

61. 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).

62. Mikolov, T., Deoras, A., Povey, D., Burget, L. & Cernocky, J. H. Strategies for Training Large Scale Neural Network Language Models. in *IEEE Workshop on Automatic Speech Recognition and Understanding* (2011).

63. Swinney, D. A. Lexical Access during Sentence Comprehension: (Re)Consideration of Context Effects. *J. Verbal Learn. Behav.* **18,** 645–659 (1979).

64. Petten, C. Van & Luka, B. J. Prediction during language comprehension: Benefits , costs , and ERP components. *Int. J. Psychophysiol.* **83,** 176–190 (2012).

65. Thornhill, D. E. & Petten, C. Van. Lexical versus conceptual anticipation during sentence processing : Frontal positivity and N400 ERP components. *Int. J. Psychophysiol.* **83,** 382–392 (2012).

66. Delong, K. A., Quante, L. & Kutas, M. Neuropsychologia Predictability , plausibility , and two late ERP positivities during written sentence comprehension. *Neuropsychologia* **61,** 150–162 (2014).

67. Osterhout, L. & Holcomb, P. J. Event-Related Brain Potentials Elicited by Syntactic Anomaly. *J. Mem. Lang.* **31,** 785–806 (1992).

68. Brouwer, H. & Hoeks, J. C. J. A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. *Front. Hum. Neurosci.* **7,** 758 (2013).

69. Coulson, S., King, J. W. & Kutas, M. Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Lang. Cogn. Process.* **13,** 21–58 (1998).

70. Sassenhagen, J., Schlesewsky, M. & Bornkessel-Schlesewsky, I. The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following

linguistically deviant material is reaction time aligned. *Brain Lang.* **137,** 29–39 (2014).

71. Polich, J. Updating P300 : An integrative theory of P3a and P3b. *Clin. Neurophysiol.* **118,** 2128–2148 (2007).

72. Schacht, A., Sommer, W., Shmuilovich, O., Casado Martinez, P. & Martin-Loeches, M. Differential Task Effects on N400 and P600 Elicited by Semantic and Syntactic Violations. *PLoS One* **9,** 1–7 (2014).

73. Luck, S. J., Vogel, E. K. & Shapiro, K. L. Word meanings can be accessed but not reported during the attentional blink. *Nature* **383,** 616–618 (1996).

74. Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E. & Perry, N. W. Brain potentials related to stages of sentence verification. *Psychophysiology* **20,** 400–409 (1983).

75. Staab, J. *et al.* Negation Processing in Context Is Not (Always) Delayed. **20,** (2008).

76. Nieuwland, M. S. & Kuperberg, G. R. When the Truth Is Not Too Hard to Handle. *Psychol. Sci.* **19,** 1213–1218 (2008).

77. 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).

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

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

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

81. 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

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

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

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

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

86. Gleitman, L. R., January, D., Nappa, R. & Trueswell, J. C. On the give and take between event apprehension and utterance formulation. *Mem. Lang.* **57,** 544–569 (2007).

87. Altmann, G. T. M. & Kamide, Y. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* **73,** 247–264 (1999).

88. Kamide, Y., Altmann, G. T. M. & Haywood, S. L. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Mem.*

Lang. **49,** 133–156 (2003).

89. 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).

90. Elman, J. L. Finding Structure in Time. *Cogn. Sci.* **14,** 179–211 (1990).

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/* … the eggs would *eat/* … the boys would *plant*	cong. =< rev. anom. < incong.	Kuperberg, Sitnikova, Caplan, & Holcomb (2003) , Hoeks et al. (2004)
Word order violation	She is very satisfied with the ironed neatly linen	no effect	Hagoort & Brown (2000)
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’)