nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0406-4

Modelling the N400 brain potential as change in a probabilistic representation of meaning

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

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. However, the underlying process remains incompletely understood and actively debated. Here we present a computationally explicit account of this process and the emerging representation of sentence meaning. We simulate N400 amplitudes as the change induced by an incoming stimulus in an implicit and probabilistic representation of meaning captured by the hidden unit activation pattern in a neural network model of sentence comprehension, and we propose that the process underlying the N400 also drives implicit learning in the network. The model provides a unified account of 16 distinct findings from the N400 literature and connects human language comprehension with recent deep learning approaches to language processing.

The N400 component of the event-related brain potential (ERP) has received a lot of attention, as it promises to shed light on the brain basis of meaning processing. The N400 is a negative brain potential recorded over centro-parietal areas 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 coffee with cream and…' the anomalous word 'dog' produces a larger N400 than the congruent word 'sugar'[1]. The N400 has since been used as a dependent variable in over 1,000 studies[2]. However, despite the large amount of data on the N400, its functional basis continues to be debated: various verbal descriptive theories have been proposed[3–7], but their capacity to capture all the relevant data is difficult to determine unambiguously due to the lack of implementation, and none has yet offered a generally accepted account[2].

Here, we provide both support for and formalization of the view that the N400 reflects the input-driven update of a representation of sentence meaning—one that implicitly and probabilistically represents all aspects of meaning as it evolves in real time during comprehension[2]. We do so by presenting an explicit computational model of this process. The model is trained and tested using materials generated from a simplified artificial microworld in which we can manipulate variables that have been shown to affect the N400, allowing us to explore how these factors affect processing. The use of synthetic materials prevents us from simulating N400s to specific sentences used in empirical experiments. However, an artificial environment provides more transparency concerning the factors influencing model behaviour than would be afforded by a naturalistic corpus.

The model does not exactly correspond to any existing account of the N400, as it implements a distinct perspective on language comprehension. Existing accounts are often grounded (at least partly) in modes of theorizing originating from the 1950s[8], in which symbolic representations 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 sentence meaning[9–11]. Although perspectives on language processing have evolved in various ways, many researchers maintain the notion that word meanings are retrieved from memory before being assigned to roles in a compositional representation.

Our model, called the sentence gestalt (SG) model[12,13], provides an alternative to this mode of theorizing. It offers a functional-level characterization of language understanding in which each word in a sentence provides clues that constrain the formation of an implicit probabilistic representation of the event described by the sentence. Earlier work established that the model could capture several core aspects of language comprehension[13]. The current work extending the model to address N400s at this functional level complements efforts to model neurophysiological details underlying the N400[14–16].

The design of the SG model (Fig. 1) reflects the principle that listeners continually update their probabilistic representation of the described event as each incoming word of a sentence is presented. The representation corresponds to an internal (hidden layer) activation state called the sentence gestalt that depends on connection-based knowledge in the 'update' part of the network (Fig. 1). The SG activation pattern can guide responses to potential queries about the event described by the sentence (see the 'Implicit probabilistic theory of meaning' section in the Supplementary Discussion). The model is trained with sentences and queries about the events the sentences describe, so that it can provide responses to such queries. Although we focus on a simple microworld of events and sentences that describe them, the model exemplifies a wider conception of a neural activation state that represents a person's subjective understanding of a broad range of situations and of the kinds of inputs that can update this understanding. The input could be in the form of language expressing states of affairs, or even non-declarative language such as questions (Supplementary Discussion). Although we focus on linguistic input here, the input guiding the formation of this representation could also come from witnessing events directly, from pictures or sounds, or from any combination of linguistic and other forms of input.

The magnitude of the activation update produced by each successive word of a sentence corresponds to the change in the model's probabilistic 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

Department of Psychology, Stanford University, Stanford, CA, USA. *e-mail: milena.rabovsky@gmail.com; mcclelland@stanford.edu

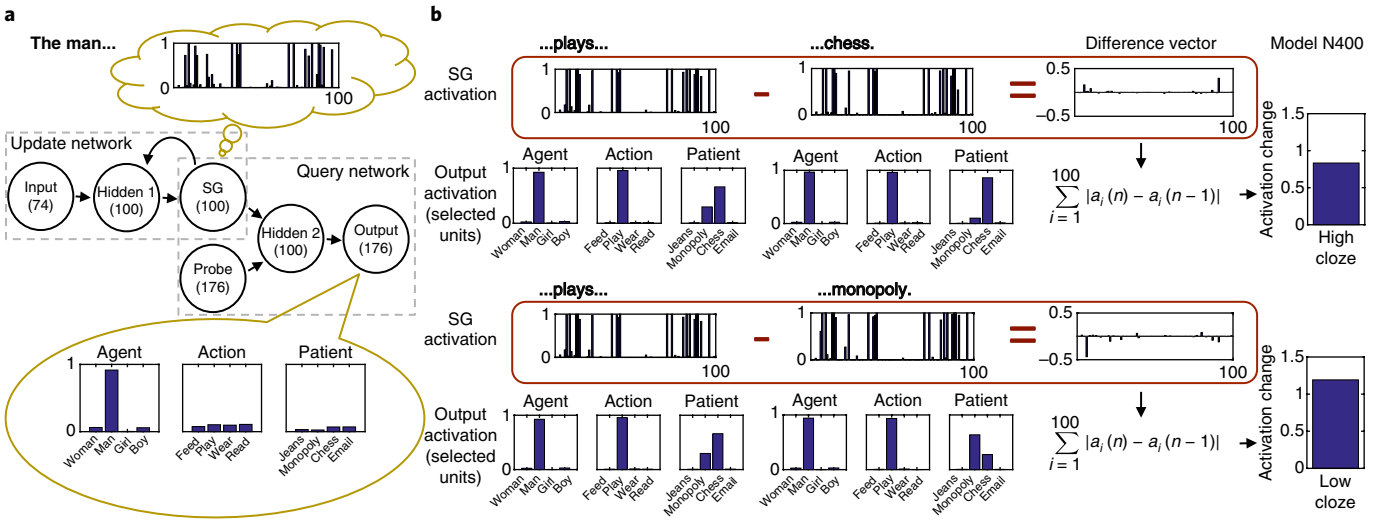

Fig. 1 | The sentence gestalt (SG) model architecture, processing a sentence with a high- or low-cloze probability ending, and the model's N400 correlate. a, The model consists of an update network and a query network. Circles represent layers of units (and number of units in each layer). Arrows represent 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 SG activation to produce the updated SG pattern corresponding to the updated implicit representation of the described event. During training, after each presented word, the model is probed concerning all aspects of the described event (for example: agent, 'man'; action, 'play', and so on) in the query network. Here, the activation from the probe layer combines via layer 'hidden 2' with the current SG pattern to produce output activations. Selected output units activated in response to the agent, action and patient probes are shown; each query response includes a distinguishing feature (such as 'man' or 'woman', as shown) as well as other features (such as 'person' or 'adult', not shown) that capture semantic similarities among event participants (Supplementary Table 1). After presentation of 'The man', the SG representation (thought bubble, top) supports activation of the correct features when probed for the agent and estimates the probabilities of action and patient features. **b**, After the word 'plays', 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, based on its experience (the man plays chess more often than monopoly). If the next word is 'chess' (top), the change in SG activation (summed magnitudes of changes in 'difference vector') is smaller than if the next word is 'monopoly' (bottom). The change, called the semantic update, is the proposed N400 correlate (right), which is larger for the less probable ending (in this case, 'monopoly', bottom).

absolute value of the change in each unit's activation that this word produces. 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 $n-1$:

$$\mathrm{N400}_n = \mathrm{SU}_n = \sum_i |a_i(w_n) - a_i(w_{n-1})|$$

This measure can be related formally to a Bayesian measure of surprise[17] and to the signals that govern learning in the network (see below and Supplementary Discussion).

How the SU captures the N400 is best illustrated with an example: 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 word 'sugar' is presented, resulting in little change in activation, and thus a small N400. In contrast, the representation will change much more if 'dog' is presented instead, corresponding to a much larger change in subjective probabilities of event characteristics, reflected in a larger change in activation and thus a larger N400.

Distinctive features of the sentence gestalt model

Several aspects of the model's design and behaviour are worth exploring to understand how it accounts for the empirical findings. First, the model forms a representation of the situation or event described by the sentence, rather than a representation of the sentence itself. This contrasts with models of language processing that focus primarily on updating linguistic expectations, including expectations about specific words or structural relationships[10,11]. Furthermore, unlike most other models, the SG model does not contain separate modules for lexical access or syntactic parsing—instead it simply maps from word forms to an implicit probabilistic representation of sentence meaning.

We make no stipulations of the form or structure of the model's internal representations; rather, these representations are shaped by the statistics of its experiences[18,19]. To train the model, we need a way of providing it with information about the event described by the sentence. Similar to the original implementation, events are described in terms of an action, a location, a situation (such as 'at breakfast'), an agent, and the object or patient to which the action is applied. Critically, the event description is not the model's internal representation, but is instead a characterization of those aspects of the event that the representation should be capable of describing if probed. In this way our model is similar to contemporary deep learning models such as Google's neural machine translation (GNMT) system[20], which also makes no stipulations of the form or structure of the internal representation generated from an input sentence; instead, the representation is shaped by learning to predict the translation of a sentence from one language in other languages. The relative success of the GNMT 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].

Second, learning takes place in the model over an extended time course loosely corresponding to human development, based on the gradual accumulation of experience about events and the sentences that describe them. Thus, the SU occurring on presentation

Table 1 | Overview of simulated N400 effects

Simulated effects	Example	N400 data	Ref. no.
Basic effects			
(1) Semantic incongruity	I take coffee with cream and **sugar/dog**.	Cong. < incong.	1
(2) Cloze probability	Don't touch the wet **paint/dog**.	High < low	25
(3) Position in sentence		Late < early	26
(4) 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.	29
(5) Lexical frequency		High < low	31
(6) Semantic priming	Sofa – bed	Related < unrelated	32
(7) Associative priming	Wind – mill	Related < unrelated	32
(8) Repetition priming		Rep. < unrelated	33
Specificity of the N400 effect			
(9–11) Reversal anomalies	(1) For breakfast the eggs would only **eat**...	Cong. ≤ reversal < incong.	34
	(2) The javelin has the athletes **thrown**. (In Dutch)		66
	(3) The fox that on the poacher **hunted**... (In Dutch)		37
(12) Word order violation	She is very satisfied with the ironed neatly linen.	No effect	38
(13) Constraint for unexpected endings	Joy was too frightened to **look** (low constraint). The children went out to **look** (high constraint).	No effect	39
Development and learning			
(14) Age		Babies: less compr. < more compr. Later: young > old	40–42
(15) Priming during near chance second language performance	Chien – chat	Related < unrelated	44
(16) Repetition × incongruity		Cong. (|nonrep. – rep.|) < incong. (|nonrep. – rep.|)	48

Cong., congruent; incong., incongruent; cat. rel., categorically related; reversal, reversal anomaly; rep., repeated; compr., comprehension; nonrep., nonrepeated.

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 N400s as a function of experience.

Third, the model responds to words presented to it, independently of whether they form sentences, as implemented in the update part of the network (Fig. 1). This allows the model to address N400s evoked by words presented in pairs or in isolation. We view this as a largely automatic process, proceeding independently of the intention of the listener. Whether they are in sentences or not, the SG activity produced by words will reflect aspects of events in which they occur, in line with embodied approaches to the representation of meaning[23,24]. The explicit computation of responses to queries about events is used during training to allow the model to learn to map from sentences to meaning, but this process is not thought of as contributing to the N400, and would not ordinarily occur during an N400 experiment, when no external event information is available.

Finally, we do not see the process reflected in the N400 as the only process that contributes to language understanding. Other processes, reflected in other ERP components, may come into play in understanding sentences describing implausible events or sentences with unusual structure, and these processes may result in changes to the meaning representation that is ultimately derived from a linguistic input. In the 'Discussion' section, we consider how the formation of an initial, implicit representation of meaning, as captured by the SG model, might fit into a broader picture of language understanding.

Results

We report 16 simulations of well-established N400 effects, chosen to illustrate how the model can address a broad range of empirical findings taken as supporting diverse descriptive theories of the N400's functional basis (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.

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

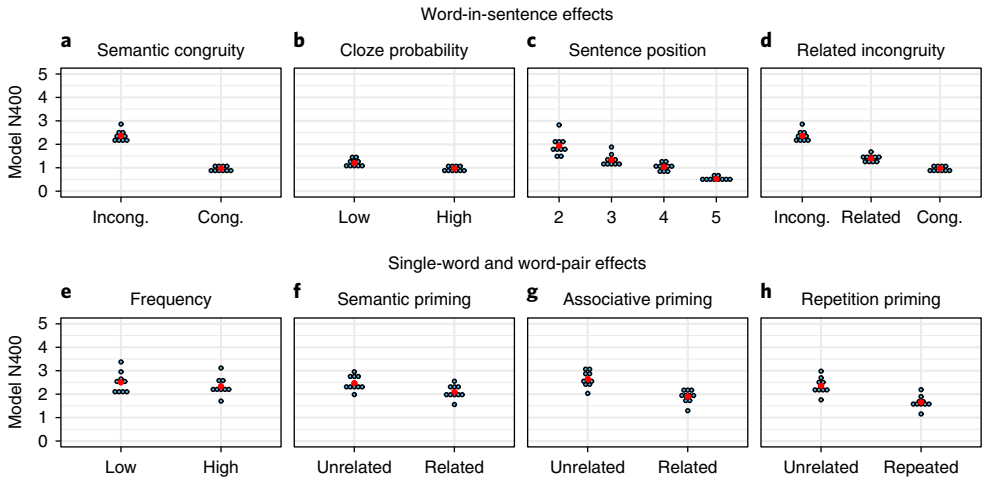

Fig. 2 | Simulation results for the basic effects. a–h, The model's N400 correlate. Blue dots represent results for independent runs of the model ($n=10$ models), averaged across items per condition (n values for items given below); red dots represent condition means; ± standard error of the mean (SEM) would be represented by red error bars, but in this figure the error bars do not exceed the area of the red dot. Cong., congruent; incong., incongruent. Statistical results for **a–h** are shown below: t_1 from the model analyses (item analyses are reported in Supplementary Fig. 1), Cohen's d (note that effect sizes might be larger in simulations than in empirical experiments due to the noise in EEG signals) and 95% confidence interval for the condition difference, CI). **a**, Semantic incongruity ($n=10$ items per condition): $t_1(9)=25.00$, $P<0.001$, $d=7.91$, 95% CI (1.26, 1.51). **b**, Cloze probability ($n=10$ items): $t_1(9)=8.56$, $P<0.001$, $d=2.71$, 95% CI (0.18, 0.30). **c**, Position in sentence ($n=12$ items): $t_1(9)=8.17$, $P<0.001$, $d=2.58$, 95% CI (0.43, 0.76) from second to third sentence position; $t_1(9)=4.73$, $P=0.003$, $d=1.50$, 95% CI (0.16, 0.44) from third to fourth position; and $t_1(9)=17.15$, $P<0.001$, $d=5.42$, 95% CI (0.44, 0.58) from fourth to fifth position. **d**, Categorically related incongruities ($n=10$ items) were larger than congruent, $t_1(9)=10.63$, $P<0.001$, $d=3.36$, 95% CI (0.33, 0.51), and smaller than incongruent continuations, $t_1(9)=14.69$, $P<0.001$, $d=4.64$, 95% CI (0.82, 1.11). **e**, Lexical frequency ($n=14$ items): $t_1(9)=3.13$, $P=0.012$, $d=0.99$, 95% CI (0.05, 0.31). **f**, Semantic priming ($n=10$ items): $t_1(9)=14.55$, $P<0.001$, $d=4.60$, 95% CI (0.32, 0.44). **g**, Associative priming ($n=10$ items): $t_1(9)=14.75$, $P<0.001$, $d=4.67$, 95% CI (0.63, 0.86). **h**, Immediate repetition priming ($n=10$ items): $t_1(9)=16.0$, $P<0.001$, $d=5.07$, 95% CI (0.60, 0.80). Details on the statistics including normality tests are reported in Supplementary Methods 4.

specific word in offline sentence completion tasks), as in the example 'Don't touch the wet dog (low cloze)/paint (high cloze)'[25]. This is also captured by the model: it exhibits a larger SU for sentence endings presented with a low (0.3) as compared to a high probability (0.7) during training (simulation 2; Fig. 2b, Supplementary Fig. 1b and Supplementary Note 1). The graded character of the underlying process is further supported by the finding that N400s gradually decrease across the sequence of words in normal congruent sentences[26]. The SU in the model correspondingly shows a gradual decrease across successive words in sentences (simulation 3; Fig. 2c and Supplementary Fig. 1c).

Expectancy for words versus 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 (word surprisal[27]) 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 to predict the next word based on preceding context[28]. However, there is evidence that N400s may not be a function of word probabilities per se but rather of probabilities of aspects of meaning signalled by words: N400s are smaller for incongruent completions that are closer semantically to the correct completion than those that are semantically more distant. Consider the following sentence: 'They wanted to make the hotel look more like a tropical resort. So, along the driveway they planted rows of…'. The N400 effect 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 from another basic level category than the congruent completion)"[29]. Our model captures these results: we compared sentence completions that were presented with a high probability during training and two types of never-presented completions. The SU was lowest for high-probability completions, as expected; crucially, among never-presented completions, SU was smaller for those that shared semantic features with high-probability completions compared to those that did not (simulation 4; Fig. 2d and Supplementary Fig. 1d).

Semantic integration versus lexical access. The sentence-level effects considered above have often been taken to indicate that N400s reflect the effort to integrate an incoming word into the preceding context[7,30]. However, a sentence context is not actually needed: N400 effects can also be obtained for words presented in pairs or in isolation. Specifically, N400s are smaller for isolated words with high rather than low lexical frequency[31]; for words (for example 'bed') presented after categorically related primes ('sofa') or associatively related primes ('sleep') as compared to unrelated primes[32]; and for immediate repetition of words compared to the same words following unrelated primes[33]. Such N400 effects outside of a sentence context, especially the influences of repetition and lexical frequency, have led some researchers to suggest that N400s do not reflect the formation of a representation of sentence meaning but only access to the current word's meaning[3,14]. As previously noted, the SG pattern probabilistically represents the meaning of a sentence if one is presented, but the model also processes words presented singly or in pairs. Unlike traditional models, there is no separate system for accessing meanings of words. Instead the model contains a single system that processes words and word sequences, whether or not they form a meaningful sentence.

The model captures all of the above-mentioned effects: first, the SU was smaller for isolated words that occurred relatively frequently during training (simulation 5; Fig. 2e and Supplementary Fig. 1e). Furthermore, the SU was smaller for words presented after words from the same semantic category compared to words from a different category (simulation 6; Fig. 2f and Supplementary Fig. 1f), and smaller for words presented after associatively related words ('chess'

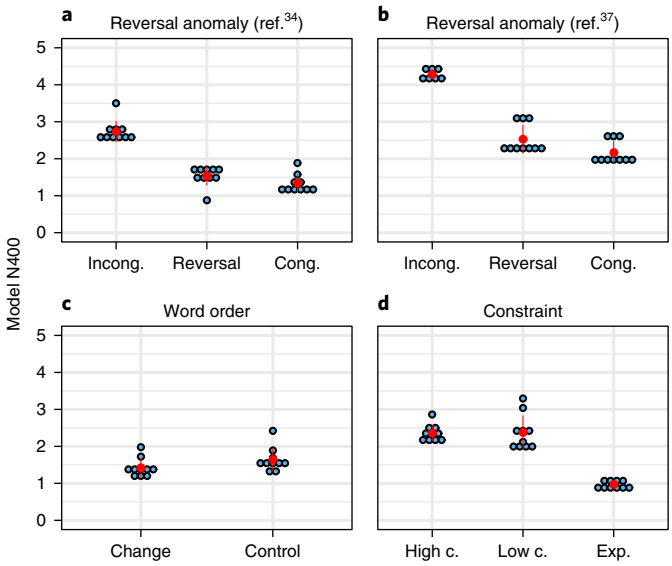

Fig. 3 | Simulation results concerning the specificity of the N400 effect. Blue dots represent results for independent runs of the model ($n=10$ models), averaged across items per condition; red dots represent condition means, red error bars represent ±SEM (see Supplementary Fig. 2 for item-based analyses). Incong., incongruent; reversal, reversal anomaly; cong., congruent; high c., unexpected high constraint; low c., unexpected low constraint; exp., expected. **a**, Reversal anomaly in standard model[34] ($n=8$ items): $t_1(9)=2.09$, $P=0.199$, $d=0.66$, 95% CI (0.02, 0.41) for comparison between congruent and reversal; $t_1(9)=10.66$, $P<0.001$, $d=3.37$, 95% CI (0.95, 1.46) for comparison between reversal and incongruent; and $t_1(9)=28.39$, $P<0.001$, $d=8.98$, 95% CI (1.29, 1.51) for comparison between congruent and incongruent. **b**, Reversal anomaly where both participants can be agents[37] ($n=8$ items). These results are from a model trained on a different environment (see the main text for details), explaining the difference in SU in the baseline (congruent) condition. Again, SU in the reversal anomaly is only slightly increased as compared to the congruent condition, while being considerably larger in the incongruent condition. Congruent versus reversal: $t_1(9)=13.25$, $P<0.001$, $d=4.19$, 95% CI (0.30, 0.42); congruent versus incongruent: $t_1(9)=55.10$, $P<0.001$, $d=17.41$, 95% CI (2.26, 2.45); reversal versus incongruent: $t_1(9)=52.21$, $P<0.001$, $d=16.51$, 95% CI (1.90, 2.08). **c**, Change in word order ($n=10$ items). The SU was slightly larger for normal versus changed order; significant only over models, $t_1(9)=5.94$, $P<0.001$, $d=1.88$, 95% CI (0.14, 0.31). **d**, Constraint for unexpected endings ($n=10$ items). The SU did not differ between unexpected high versus low constraint, $t_1(9)=0.13$, $P=0.90$, $d=0.04$, 95% CI (−0.24, 0.27). For expected endings it was lower than for unexpected high constraint, $t_1(9)=25.00$, $P<0.001$, $d=7.91$, 95% CI (1.26, 1.52), and unexpected low constraint, $t_1(9)=10.21$, $P<0.001$, $d=3.23$, 95% CI (1.09, 1.72).

following 'play') as compared to unrelated words ('chess' following 'eat') (simulation 7; Fig. 2g and Supplementary Fig. 1g). Finally, the SU was smaller for the second presentation of a repeated word compared to a word presented after an unrelated word (simulation 8; Fig. 2h, Supplementary Fig. 1h).

Specificity of the N400 effect. *Reversal anomalies and the N400.* A finding that has puzzled the N400 community is the lack of a robust N400 effect in 'reversal anomaly' sentences. Only a very small N400 increase occurs in sentences such as 'For breakfast the eggs would only eat…' when compared with corresponding congruent sentences such as 'For breakfast, the boys would only eat…'[34]. The small N400 effect here is typically accompanied by an increase in the P600, a subsequent positive potential. In contrast, N400 but not P600 amplitudes are considerably larger in sentence variations such as 'For breakfast the boys would only bury…'[34].

This pattern of results has sparked considerable theoretical uncertainty. Many researchers have taken these findings to indicate that the word 'eggs' in the given context is easily integrated into a representation of sentence meaning because 'eggs' is (at least temporarily) interpreted as specifying the object eaten rather than the agent of eating. Such a situation has been called a temporary 'semantic illusion'[35]. Our account is partly in line with this view, although we describe such a state of mind as an 'event probability-based interpretation' to avoid the implication that syntax must always be the definitive cue when syntax and other considerations conflict. Others[36] have taken this finding to indicate that the N400 is not related to sentence meaning but instead reflects retrieval of word meaning. The idea is that the retrieval of the meaning of 'eat' is facilitated by priming from 'breakfast' and 'eggs', whereas 'bury' would not be facilitated by prior context. On this view, understanding sentence meaning is associated with the P600 rather than the N400.

We address this controversy by showing that our model captures the small N400 effect in reversal anomalies. In the first relevant simulation, in line with N400s, the model exhibited only a slight increase in SU for reversal anomalies ('At breakfast, the eggs eat…') as compared to congruent continuations ('At breakfast, the man eats…), and a substantial increase for incongruent continuations ('At breakfast, the man plants…') (simulation 9; Fig. 3a and Supplementary Fig. 2a). The query network's responses to relevant probes suggests that the model does indeed maintain an event probability-based interpretation, in that it continues to favour the eggs as the patient instead of the agent of eating even after the word 'eat' is presented (Supplementary Fig. 3).

The second simulation addresses the small N400 effect in reversal anomalies in which both participants are animate beings that can occur as agents (this was not the case in the first simulation and another (simulation 10; Supplementary Fig. 4) described in Supplementary Methods 1). Consider these materials[37]: 'De vos die op de stroper joeg…' ('The fox who hunted the poacher…') and 'De zieke die in de chirurg sneede…' ('The patient who cut into the surgeon…'). Here, both event participants are animate, yet the syntactically supported interpretations are inconsistent with event probabilities. (We use the phrase 'event probabilities' to refer to the probability distribution of role fillers in events consistent with the words so far encountered, independent of the order of the words. For example, at the second noun in 'the poacher on the fox' and 'the fox on the poacher', the words so far encountered are the same, and so event probabilities would be the same as well.) Both participants can be agents in events involving the other participant (fox could watch poacher, and patient could stand in front of surgeon), and both can engage in the relevant action (hunt something, or cut into something). What makes these cases anomalous is that in hunting events involving poachers and foxes, it is always the poachers that hunt the foxes; and in events involving surgeons and patients where one is cutting into the other, it is always the surgeons that cut into the patients.

To address such cases, we conducted a simulation focusing on the experiment that used the cited examples (among others)[37]. The experiment was done in Dutch; this is critical because it means that both nouns are presented before the verb. We therefore trained an additional model with Dutch word order, using event scenarios set up to align with the materials used in the target experiment (Supplementary Methods 2 and Supplementary Fig. 5). Using sentences from these scenarios, the model again successfully captures the N400 in reversal anomalies. It exhibited only a very slight increase in SU for reversal anomalies ('The fox on the poacher hunted') compared to congruent sentences ('The poacher on the fox hunted') and a substantial increase for incongruent

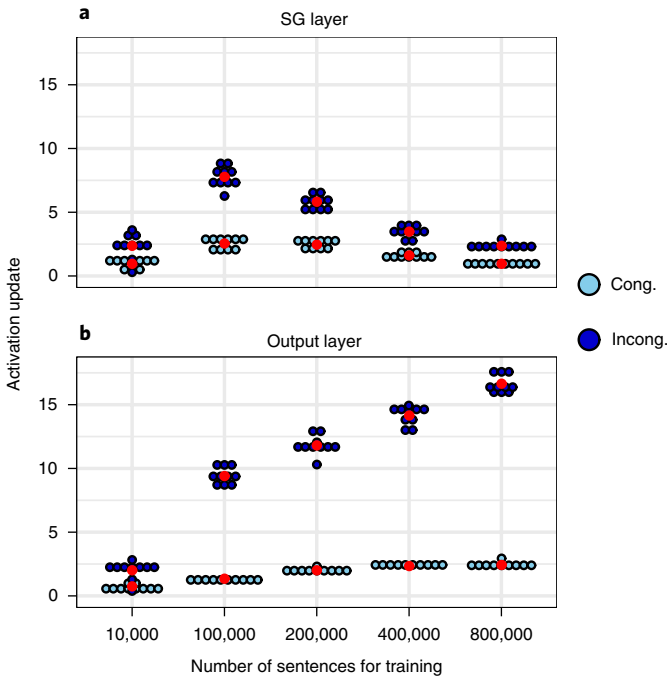

Fig. 4 | Development across training. Semantic incongruity effects as a function of the number of sentences the model has been exposed to. **a**, Semantic update at the model's hidden SG layer shows at first an increase and later a decrease with additional training, in line with the developmental trajectory of the N400. Each light blue or dark 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 (see Supplementary Fig. 7 for item-based analyses). The size of the effect (the numerical difference between the congruent and incongruent condition) differed between all subsequent time points: $t_1(9) = 17.02$, $P < 0.001$, $d = 5.38$, 95% CI (3.28, 4.29) between 10,000 and 100,000 sentences; $t_1(9) = 7.80$, $P < 0.001$, $d = 2.47$, 95% CI (1.33, 2.41) between 100,000 and 200,000 sentences; $t_1(9) = 14.69$, $P < 0.001$, $d = 4.65$, 95% CI (1.24, 1.69) between 200,000 and 400,000 sentences; $t_1(9) = 7.70$, $P < 0.001$, $d = 2.43$, 95% CI (0.34, 0.62) between 400,000 and 800,000 sentences. **b**, 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.

continuations ('The poacher on the fox planted'; simulation 11; Fig. 3b and Supplementary Fig. 2b).

To understand why the model does not exhibit a substantially larger SU in the role-reversed sentences compared to controls, we examined the network's responses to relevant probes. While the model's interpretation of the congruent sentences was unambiguous, it exhibited uncertainty in its role assignments when processing reversal anomalies, due to conflicting constraints imposed by word order and event probabilities. This conflict was not reflected in a large SU at the verb because it was already present at the second noun and was not resolved by the verb (see Supplementary Note 2 and Supplementary Fig. 6).

In summary, the simulations show that the small N400 effect in reversal anomalies is consistent with the view that the N400 reflects the updating of an implicit representation of sentence meaning as implemented in the SG model. The model is partly in line with previous accounts favouring a role for plausibility constraints in sentence processing[35]. However, in our model, the initial heuristic comprehension process underlying N400s is not purely based on event probabilities. Instead, the model is sensitive to both event probabilities and syntactic constraints, and from this perspective the small N400 effect in reversal anomalies does not necessarily reflect a clear-cut event probability-based interpretation (of, for instance, the poacher hunting the fox). Instead, the finding may reflect a state of unresolved conflict between different cues. Other processes, possibly associated with the P600, may resolve the conflict between competing interpretations in such situations (see Discussion).

Specificity of the N400 to violations of semantic rather than syntactic factors. N400s are not influenced by syntactic factors such as violations of word order (such as 'The girl is very satisfied with the ironed neatly linen'), which instead elicit P600 effects[38]. Because the model is representing the event described by the sentence, and this event is not necessarily affected by a change in word order, the model is likewise insensitive to such violations. To demonstrate this, we considered the model's response to changes in word order ('On Sunday, the man *the robin* feeds' versus 'On Sunday, the man *feeds* the robin'), examining the SU at the highlighted position. If anything, the SU was slightly larger in the condition with normal compared to changed word order (simulation 12; Fig. 3c and Supplementary Fig. 2c), because changes in word order also entail changes in the amount of information a word provides about the described event, and the amount of semantic information was on average slightly higher in sentences with normal compared to changed word order (see Methods).

No influence of constraint for unexpected endings. The model also captures the finding that the N400 does not depend on the prior establishment of a specific expectation[39]. That is, the N400 for an unpredictable word is equally large independent of whether the word is unpredictable because the context does not predict any specific word (e.g., "Joy was too frightened to *look*.") or because the context predicts a specific word different from the one presented (e.g., "The children went outside to *look*", where *play* would be expected). Correspondingly, in the model SU was equally large for words that are unexpected because the context is unconstraining (e.g., "The man likes the *email*.") as for words that are unexpected because they violate specific expectations (e.g., "The man eats the *email*." Simulation 13; Fig. 3d, Supplementary Fig. 2d). This finding highlights the fact that the N400, like the SU in the model, corresponds to the amount of unexpected semantic information (in the sense of Bayesian surprise) and does not constitute a violation signal per se.

Development and learning. In all of the simulations above, it would have been possible to model the phenomena by treating the N400 as a reflection of change in explicit 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 SU (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 SU 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 Supplementary Discussion for details of these measures). We then consider how the implicit SU can drive connection-based learning in the update network, accounting for a final pattern of empirical findings.

Development. N400s change with increasing language experience and over development. The examination of N400 effects in different age groups has shown that N400 effects increase with comprehension skills very early in life[40] but later decrease with age[41,42].

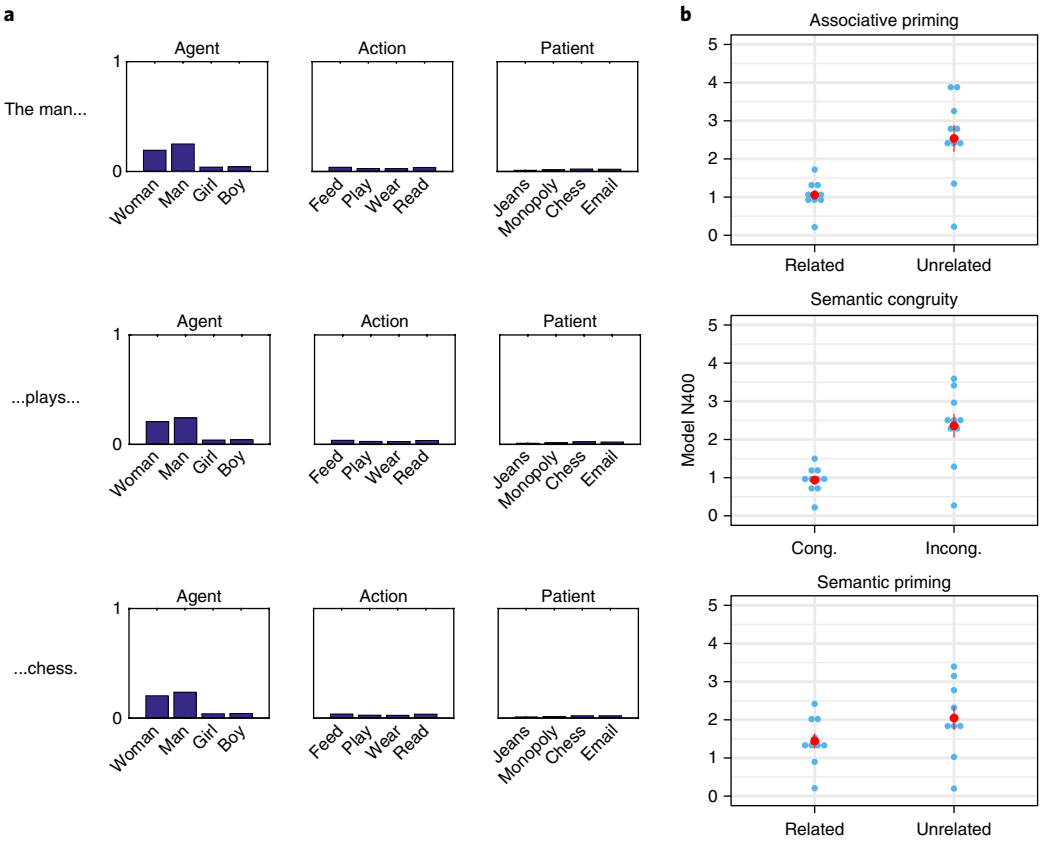

Fig. 5 | Comprehension performance and semantic update effects at a very early stage in training. a, Activation of selected output units while the model is presented with the sentence 'The man plays chess'. It can be seen that the model fails to activate the corresponding units at the output layer. The 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 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 ($t_1(9) = 6.12$, $P < 0.001$, $d = 1.94$, 95% CI (0.93, 2.03), top), semantic congruity in sentences ($t_1(9) = 6.85$, $P < 0.001$, $d = 2.16$, 95% CI (0.95, 1.90), middle) and semantic priming ($t_1(9) = 5.39$, $P < 0.001$, $d = 1.70$, 95% CI (0.35, 0.85), 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 (see Supplementary Fig. 8 for item-based analyses). Cong., congruent; incong., incongruent.

A comparison of the effect of semantic congruity on SU at different points in training shows a developmental pattern consistent with these findings (simulation 14; Fig. 4a and Supplementary Fig. 7a): 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. 4b and Supplementary Fig. 7b). The activation pattern at the output layer reflects explicit estimates of semantic feature probabilities in that units at the output layer explicitly represent semantic features (such as 'can grow' and 'green'), 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 (for example, activation of 0.7 when the conditional probability of the feature is 0.7). Thus, in the trained model, changes in output activation induced by an incoming word approximate changes in explicit estimates of feature probabilities induced by that word. The continuing increase of the congruity effect across training (Fig. 4b) shows that it is not the changes in the model's explicit estimates of semantic feature probabilities that pattern with the developmental trajectory of the N400 effect. Instead, it is the change in the hidden SG layer activation that corresponds to this developmental trajectory (Fig. 4a).

The reversed pattern for changes in hidden and output activations is possible because, as noted above, the SG activation does not explicitly represent the probabilities of semantic features—instead, it provides a basis (together with 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 labour from activation to connection weights is interesting in that it might underlie the common finding that experience often leads to decreased neural activity in parallel with increased speed and accuracy of task performance[43].

Early sensitivity to a new language. A study of human second language learning showed robust influences of semantic priming on N400s while overt lexical decision performance in the newly trained language was still near chance[44]. Although second language learning is beyond the scope of the present work, we observe a similar pattern at a very early stage in our model's training (Fig. 5a). At this early stage, overt estimates of feature probabilities were only weakly modulated by the words presented, but the SU was significantly influenced by semantic priming, associative priming and semantic congruity in sentences (simulation 15; Fig. 5b and Supplementary Fig. 8). Similarly to humans, then, our model can exhibit N400-like effects before overt behaviour robustly reflects learning.

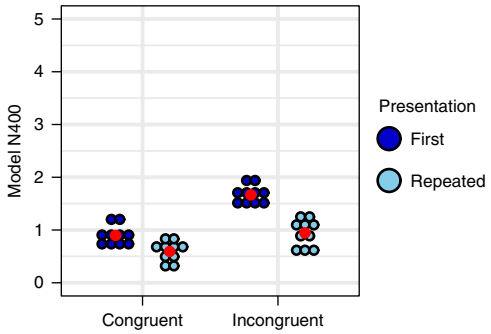

Fig. 6 | Simulation of the interaction between delayed repetition and semantic incongruity. Each dark blue or light blue dot represents the results for one independent run of the model, averaged across items per condition; the red dots represent the means for each condition (see Supplementary Fig. 9 for item-based analyses). There were significant main effects of congruity, $F_1(1,9) = 214.13$, $P < .001$, $\eta_p^2 = 0.960$, and repetition, $F_1(1,9) = 48.47$, $P < 0.001$, $\eta_p^2 = 0.843$, and a significant interaction between both factors, $F_1(1,9) = 83.30$, $P < 0.001$, $\eta_p^2 = 0.902$. Post-hoc comparisons showed that even though the repetition effect was larger for incongruent as compared to congruent sentence completions, that is, incongruent (first – repetition) > congruent (first – repetition), $t_1(9) = 9.13$, $P < 0.001$, $d = 2.89$, 95% CI (0.32, 0.53), it was significant in both conditions, $t_1(9) = 4.21$, $P = 0.0046$, $d = 1.33$, 95% CI (0.14, 0.46) for the congruent completions, and $t_1(9) = 8.78$, $P < 0.001$, $d = 2.78$, 95% CI (0.54, 0.91) for the incongruent completions.

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 words, and the updated estimate based on the next word in the sentence. If the estimate after word $n-1$ is viewed as a prediction, this difference can be viewed as a prediction error. It is often assumed that learning is based on such temporal difference or prediction errors[45–47] so that if N400s reflect the update of a probabilistic representation of meaning, then larger N400s should be related to greater adaptation, that is, larger adjustments to future estimates. Here we implement this idea, using the SU 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, that is, $SG_n - SG_{n-1}$ (see Methods and Supplementary Discussion for details). Importantly, this allows the model to learn from listening or reading, without a separate event description. 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 a larger N400 compared to a congruent completion, leads to a larger reduction in the N400 when the presentation is repeated after a delay, compared to the congruent continuation[48].

To simulate this pattern, we presented a set of congruent and incongruent sentences, 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 and observed a greater reduction in the N400 with repetition of incongruent compared to congruent sentence completions (simulation 16; Fig. 6 and Supplementary Fig. 9).

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[49,50]. Thus, our account predicts that in general, larger N400s should induce stronger adaptation. Although further investigation is needed, there is some evidence consistent with this prediction: larger N400s to single words during a study phase predict enhanced implicit memory (measured by stem completion without explicit memory) during testing[51].

Discussion

The N400 ERP component is widely used to investigate the neurocognitive processes underlying the processing of meaning in language. However, the component's functional basis continues to be actively debated[2]. In the simulations presented above, we have shown that an implemented computational model of language comprehension, the sentence gestalt model, provides a unified account capturing a wide range of findings (Table 1). The model treats N400 amplitudes as the change induced by an incoming word in an implicit probabilistic representation of meaning. Here we explain how the model's distinctive characteristics contribute to its ability to account for the data.

First, our model does not assume separate stages for lexical access (retrieval of word meaning) and subsequent integration of word meanings into a compositional representation. This is crucial because the two most prominent competing theories of the N400's functional basis suggest that N400s reflect either lexical access[3] or integration/unification into a compositional/combinatorial representation of sentence meaning[6,7]. In the SG model, incoming stimuli instead serve as 'cues to meaning'[52], which automatically change an activation pattern that implicitly represents estimates of conditional probabilities of all aspects of meaning of the described event. 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 doesn't represent only the currently incoming word. Instead, similar to the integration view, the resulting activation state is assumed to represent all aspects of the event described by the sentence (including—although not currently implemented—input from other modalities), but our model differs from integration accounts in avoiding a commitment to explicit compositional representation. Our perspective is in line with a recent review on the N400 that concluded that the component 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"[2]. 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. The model uses event probability together with word order to build a meaning representation instead of slotting individual word meanings into a syntactic structure. It may override syntactic conventions when event probability information is strong, or may experience uncertainty when syntactic and event probability information conflict. These aspects of the model underlie its behaviour when presented with reversal anomalies (Fig. 3a,b and Supplementary Fig. 4) or violations of word order (Fig. 3c), allowing it to explain the observed absence of N400 effects.

Second, the model's representations result from a learning process and thus depend on the statistical regularities in the model's environment as well as the amount of training. This allows the model to account for N400 effects across development (Fig. 4), including N400 effects while behavioural performance is near chance (Fig. 5) and the influence of delayed repetition on N400 congruity effects (Fig. 6).

Third, the model updates its activation after the presentation of a word, whether or not it occurs in a sentence, allowing it to capture N400 effects for single words (frequency effects; Fig. 2e) and words presented in pairs (repetition, Fig. 2h; semantic

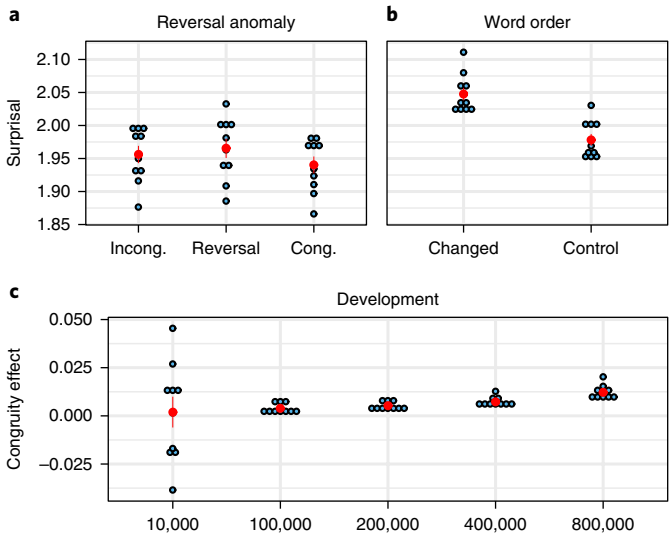

Fig. 7 | Simulation results from a simple recurrent network (SRN) model trained to predict the next word based on preceding context. Each blue dot represents the results for one independent run of the model, averaged across items per condition; red dots represent means for each condition, and red error bars represent ±SEM (see Supplementary Fig. 10 for item-based analyses). **a**, Reversal anomaly: $t_1(9) = 4.55$, $P = 0.0042$, $d = 1.44$, 95% CI (0.013, 0.038) for the comparison between congruent and reversal anomaly; $t_1(9) = 12.28$, $P < 0.001$, $d = 3.87$, 95% CI (0.013, 0.019) for the comparison between congruent and incongruent; $t_1(9) = 1.52$, $P = 0.49$, $d = 0.48$, 95% CI (−0.005, 0.024) for the comparison between incongruent and reversal anomaly. **b**, Word order: $t_1(9) = 29.78$, $P < 0.001$, $d = 9.42$, 95% CI (0.064, 0.075). **c**, Congruity effect on surprisal as a function of the number of sentences to which the model has been exposed: $t_1(9) = 0.26$, $P = 1.0$, $d = 0.082$, 95% CI (−0.015, 0.019) for the comparison between 10,000 and 100,000 sentences; $t_1(9) = 6.74$, $P < 0.001$, $d = 2.13$, 95% CI (0.0009, 0.0019) for the comparison between 100,000 and 200,000 sentences; $t_1(9) = 7.45$, $P < 0.001$, $d = 2.36$, 95% CI (0.0014, 0.0026) for the comparison between 200,000 and 400,000 sentences; $t_1(9) = 10.73$, $P < 0.001$, $d = 3.39$, 95% CI (0.0039, 0.0060) for the comparison between 400,000 and 800,000 sentences.

priming, Fig. 2f; associative priming, Fig. 2g) as well as words presented in a sentence context (semantic congruity, Fig. 2a; cloze probability, Fig. 2b; position in the sentence, Fig. 2c; semantically related incongruity, Fig. 2d; and no influence of constraint for unexpected endings, Fig. 3d).

Fourth, we propose that the N400 as captured by the model characterizes one specific aspect of language comprehension, namely the automatic stimulus-driven update of an initial implicit representation of the described event. This is in line with 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)[53,54]. The processes underlying the N400 may thus correspond to a type of language processing that has been characterized as shallow, underspecified[55], plausibility based[35] and "good enough"[56], and that may be preserved in patients with lesions to frontal cortex (specifically the left inferior prefrontal cortex, BA47)[57,58]. This region has been proposed to support control processes in comprehension that are required when processing demands are high[59,60], such as in syntactically complex sentences[57] and sentences requiring selection among competing alternative interpretations[61] (for example, reversal anomalies and garden path sentences). These aspects of language comprehension do not contribute to the N400—instead, they may be reflected in other ERP components, as discussed below.

The SG activation latently predicts the attributes of the entire event described by a sentence, capturing base-rate probabilities (before sentence processing begins) and adjusting this activation pattern as each word of the sentence is presented. In the current implementation, inputs are presented at discrete time steps corresponding to successive words, but this is a simplification for tractability. We assume that, in reality, the adjustment of semantic activation occurs continuously in time, so that the earliest arriving information about a word immediately influences the evolving SG representation[62], in line with the finding that N400 effects in spoken language comprehension often begin before the word has become acoustically unique[63,64]. It is important to note 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 so that it can respond to it with little additional change. Semantic activation changes induced by new input reflect the discrepancy between anticipated and encountered information about aspects of meaning conveyed by the sentence, and it is this discrepancy that corresponds to the learning signal driving adaptation of connection-based knowledge representations[49]. In this sense, our approach is in line with Bayesian approaches to understanding neural dynamics[46,65]. 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 increasingly efficient as learning progresses.

Recently, other studies have also linked the N400 to computational models. Most have concentrated on words presented singly or in pairs, and do not address processing in sentence contexts[14–16,49]. Two modelling studies focus on sentence processing. One of them observed a correlation between N400s and word surprisal estimated by a simple recurrent network (SRN) trained to predict the next word based on preceding context[28]. To demonstrate that an account of N400s as word surprisal fails to capture some of the phenomena our model captures, we trained an SRN on the same corpus as the SG model and repeated some critical simulations with this SRN (see Supplementary Methods 3).

First, word surprisal reflects both semantic and syntactic expectation violations, while the N400 is specific to meaning. Indeed, while SU in the SG model was insensitive to changes in word order (Fig. 3c and Supplementary Fig. 2c), surprisal in the SRN was significantly larger for syntactically anomalous compared to normal word order (Fig. 7b and Supplementary Fig. 10b). The lack of semantic specificity of word surprisal converges with the finding that the correlation between surprisal in the SRN and N400s in the above-mentioned study was observed only for content words, not for grammatical function words[28]. Furthermore, the SRN did not capture the decrease of N400 effects with age, showing instead a slight increase with training (Fig. 7c and Supplementary Fig. 10c), because surprisal is measured in terms of the estimates of word probabilities, which become sharper throughout learning. 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 larger than for incongruent continuations (Fig. 7a and Supplementary Fig. 10a)—while SU in the SG model showed only a very slight increase, in line with N400s[34] (Supplementary Fig. 11 shows 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[36]. Change in lexical activation (which is small in reversal anomalies due to priming) is linked to the N400; change in activation representing sentence meaning is assigned to the P600 component.

As discussed above, our model captures the small N400 effect in reversal anomalies because it takes both syntactic and semantic cues into account, and can favour event statistics or remain uncertain when there is a conflict between different constraints. While both the retrieval–integration model and the SG model account for the small N400 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. Further research is required to determine whether the retrieval–integration model can capture the range of N400 findings encompassed by the SG model.

The functional basis of the P600 is not addressed by our model and requires further investigation. P600 responses have been observed with a wide range of linguistic irregularities, including reversal anomalies[34,37,66], syntactic violations[38], garden path sentences[67] and pragmatic processes[68]. Some have taken these findings to suggest that the P600 might reflect combinatorial aspects of language processing, either related to syntax[38] or to semantic integration[36]. An alternative more in line with our account of the N400 links the P600 not to language processing per se, but to more conscious, deliberate and effortful aspects of processing in general. Indeed, the P600 shares properties with the oddball-sensitive P3[69,70], which has been linked to explicit surprise and working memory updating[71]. This P600-as-P3 perspective naturally explains the sensitivity of P600 effects to task demands and attention; the effect is strongly reduced or absent when there is no 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 responding to a wider range of expectation violations while the N400 is specific to meaning processing. Further, the N400 may reflect an automatic and implicit process that can result in underspecified and plausibility based representations[55,56] (see Supplementary Note 3). In contrast, the P600 may be associated with more controlled and attention-related processes, which may be affected by factors beyond those influencing N400s, and may contribute to resolving situations of cue conflict. Further research is required to better understand these issues.

Our work opens up extensive opportunities for further investigation. One key result that needs to be addressed is that N400s were observed to be unaffected by sentence truth, at least in negated statements: N400s are equally small in the false and true sentences 'A robin is not/is a bird' and equally large in the true and false sentences 'A robin is not/is a vehicle'[74]. Sentence truth is not the same as expected sentence meaning, and to understand the influence of negation on meaning expectations, the pragmatics of negation need to be taken into account (Supplementary Note 4). Studies that did this showed that N400s are indeed modulated by sentence truth[75] and plausibility[76]. Our model currently has no experience with negation and its pragmatics, but this could be incorporated in an extension. Another finding that should be addressed is that discourse meaning can influence the N400 over and above the local sentence context[77,78]. Yet another aspect to investigate is the parametric variation of corpus statistical factors contributing to the effects obtained in reversal anomalies, as the details of the model's interpretation in situations of cue conflict strongly depend on the statistics of its environment (see Supplementary Methods 2).

Finally, it remains to be explored how well the SG model can address behavioural measures of sentence processing. The beauty of ERPs is that different components index distinct aspects of processing and can thus speak to the neurocognitive reality of these aspects even though several processes might jointly influence behavioural measures. To fully address behaviour, the model will probably need to be integrated into a more complete account of the neuro-mechanistic processes taking 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 this issue and the full range of relevant neurocognitive phenomena, including other ERP components and signals that have been detected using other measurement modalities[60,79].

While extending the model will be worthwhile, it nevertheless makes a useful contribution to understanding the brain processes underlying language comprehension even in its current simple form. 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.

Methods

Here we 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.

Environment. The model environment consists of [sentence, event] pairs probabilistically generated online during training according to constraints embodied in a simple generative model (Supplementary Fig. 12a). Sentences are single-clause sentences such as 'At breakfast, the man eats eggs in the kitchen', stripped of articles as well as inflectional markers of tense, aspect and number. They 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' and '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 them. For example, the semantic features of 'daisy' are labelled '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[80,81].

For instance, all living things shared a semantic feature ('can grow'), all plants shared an additional feature ('has roots') and all flowers shared one more feature ('has petals'). The daisy then 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 (Supplementary Table 1 lists all concepts and features). Comparison of a similarity matrix of the concepts based on our handcrafted semantic representations and representations based on a principal component analysis (PCA) performed on semantic word vectors derived from co-occurrences in large text corpora[82] showed a reasonable correspondence ($r = 0.73$; Supplementary Fig. 12b), suggesting that similarities among the handcrafted 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 observes 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. This is not a principled theoretical assumption but rather just a practical consequence of the training approach. 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 from listening or reading based on the SU-driven learning rule (see Simulation 16 and Supplementary Discussion). The training procedure also implements the assumption that listeners anticipate the full meaning of each sentence as early as possible[83,84], so that the model learns to probabilistically pre-activate 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 Supplementary Fig. 12a, and then going through the following steps (please refer to Fig. 1a for the model architecture). 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) is replaced by a new activation pattern which reflects the influence of the current constituent. The SG activation pattern is then frozen while the model is probed concerning the described event in the query part of the model. Specifically, for each probe question, a unit (representing a thematic role) or units (representing feature-based representations of fillers concepts) at the probe layer are activated and feed into the hidden layer ('hidden 2'), which simultaneously 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 (that is, 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. For each response, the model's activation at the output layer is compared with the correct output, 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 SU at any given point is determined by the statistical regularities in the training set, we try to provide clarity on how the observed effects depend on the training corpus (refer to Supplementary Fig. 12a).

Basic effects. To simulate semantic incongruity (simulation 1), cloze probability (simulation 2) and categorically related semantic incongruity (simulation 4), for each condition one agent ('man') was presented once with each of the ten specific actions (excluding 'like' and 'look at'). The agent was not varied because the conditional probabilities depend very little on the agents (the only effect is that the manipulation of cloze probability is stronger for 'man' and 'woman' (0.7 versus 0.3) than for 'girl' and 'boy' (0.6 versus 0.4); Supplementary Fig. 12a). To simulate semantic incongruity, objects/patients were the high-probability objects in the congruent condition (for example, 'The man plays chess') and unrelated objects in the incongruent condition ('The man plays salmon'). To simulate cloze probability, objects were the high-probability objects in the high-cloze condition ('The man plays chess') and the low-probability objects in the low-cloze condition ('The man plays monopoly'). To simulate categorically related semantic incongruities, the congruent and incongruent conditions from the incongruity simulation were kept and there was an additional condition with objects from the same semantic category as the high and low cloze probability objects related to the action (which therefore shared semantic features at the output layer, for example 'The man plays backgammon'), but which were never presented as patients of that specific action during training (so that their conditional probability to complete the presented sentence was 0). Instead, these objects only occurred as patients of the unspecific 'like' and 'look at' actions (Supplementary Fig. 12a). For all these simulations, there were 10 items per condition; the SU was computed as the difference in SG layer activation between presentation of the action (word $n-1$) and the object (word n).

To simulate influences of a word's position in the sentence (simulation 3), we presented the longest possible sentences, that is, all sentences that occurred during training with a situation and a location. There were 12 items per condition; the SU was computed over the course of the sentences, that is, the SG difference between first and second word constitutes the SU at the second word, the SG difference between second and third word constitutes the SU at the third word, the SG difference between third and fourth word constitutes the SU at the fourth word, and the SG difference between fourth and fifth word constitutes the SU at the fifth word. See Supplementary Note 5 for the conditional probabilities of the constituents over the course of the sentence.

To simulate lexical frequency (simulation 5), the high-frequency condition comprised the high-probability objects from the 10 semantic categories, the two high-probability agents ('woman' and 'man') and two high-probability locations ('kitchen' and 'living room'). The low-frequency condition contained 10 low-probability objects, the two low-probability agents ('girl' and 'boy') and two low-probability locations ('balcony' and 'veranda'). High- and low-frequency locations were matched pairwise concerning the number and diversity of objects they are related to ('kitchen' with 'balcony' and 'living room' with 'veranda'). Before presenting the high- versus low-frequency words, we presented a blank stimulus (an input pattern consisting of all 0) to evoke the model's default activation, reflecting the encoding of base-rate probabilities in the model's connection weights. There were 14 items per condition; SU was computed as the SG difference between blank stimulus (word $n-1$) and high- or low-frequency word (word n).

To simulate semantic priming (simulation 6), for the related condition, the low- and high-probability objects of each of the 10 semantic object categories were presented subsequently as a prime-target pair (such as '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 ('sunfish chess'). To simulate associative priming (simulation 7), the related condition consisted of the 10 specific actions as primes followed by their high-probability patients as targets ('play chess'). For the unrelated condition, primes and targets were re-assigned such there was no relationship ('play eggs'). To simulate repetition priming (simulation 8), the high-probability object of each semantic category was presented twice ('chess chess'). For the unrelated condition, a high-probability object from another semantic category was presented as prime. For all priming simulations, there were 10 items per condition; SU was computed as the SG difference between prime (word $n-1$) and target (word n).

Specificity of the N400 effect. For the first simulation of reversal anomalies (simulation 9), 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 (for example, '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 ('At breakfast, the man eats…'), and for the incongruent condition, with a possible agent and an unrelated action ('At breakfast, the man plants…'). There were eight items per condition; the SU was computed as the SG difference between the second constituent, which could be an object or agent ('eggs' or 'man'; word $n-1$) and the action (word n). Supplementary Note 6 describes relevant aspects of the environment.

We also simulated another type of reversal anomaly where a relationship between two nouns is established before encountering the verb[66] (simulation 10; for example 'De speer heft de atleten geworpen', literally 'The javelin has the athletes thrown'; Supplementary Methods 1 and Supplementary Fig. 4).

Finally, we simulated reversal anomalies where both noun phrases could be agents in events (simulation 11) such as in 'De vos die op de stroper joeg' (literally 'The fox that on the poacher hunted')[37] or 'De zieken die in de chirurg sneden' (literally 'The patient that into the surgeon cut'). Both participants in such sentences can be agents, even in events involving the relevant action, and in events involving both of them and different actions. For details on the training for this simulation see Supplementary Methods 2 and Supplementary Fig. 5. For the congruent condition we presented a sentence describing the most typical event, for example 'The poacher on the fox hunted'; for the incongruent condition, we presented an unrelated action, for example 'The poacher on the fox planted'; and for the reversal anomaly, we presented the most typical action with agent and patient reversed, for example 'The fox on the poacher hunted'. There were eight items per condition; the SU was computed as the SG difference between the third word (typical agent or typical patient, for example 'poacher' or 'fox'; word $n-1$) and the action (word n). The model exhibited uncertainty in its interpretation of these reversal anomalies (Supplementary Note 2 and Supplementary Fig. 6).

To simulate influences of violations of word order[38] (simulation 12), we presented two types of word order changes for each sentence, focusing on sentences starting with a situation, because this allows to keep changes in conditional probabilities of semantic features relatively low when changing word order. For each sentence, we presented (1) a version where we changed the position of action and patient (such as 'On Sunday, the man the robin feeds' versus 'On Sunday, the man feeds the robin"; with the SU computed as the SG difference between the agent (word $n-1$) and the patient or action, respectively (word n)), and (2) a version where we changed the position of agent and action (such as 'On Sunday, feeds the man the robin' versus 'On Sunday, the man feeds the robin'; with the SU computed as the SG difference between the situation (word $n-1$) and the action or agent, respectively (word n)). Supplementary Note 7 describes the conditional probabilities of semantic features associated with the presented words in both versions. There were 16 items (8 of each type) per condition (normal versus changed word order).

To simulate influences of constraint on unexpected endings (simulation 13), we presented semantically incongruent sentences in the high constraint condition (for example, 'The man eats the email') and sentences containing an action that was presented with all 36 objects equally often in the low constraint condition (for example, 'The man likes the email'). This captures the crucial point that, in both conditions, the presented object is unexpected; in the high constraint condition, another object is highly expected and in the low constraint condition, no specific object is expected. While this slightly differs from the empirical experiment[39], where both continuations were low cloze but plausible, it is the best way to approximate the experimental situation within our environment. There were 10 items per condition; SU was computed as SG difference between the action (word $n-1$) and the object (word n).

Development and learning. To simulate the developmental trajectory of N400 effects (simulation 14) we examined the effect of semantic incongruity on SU (as described above) at different points in training, specifically after exposure to 10,000 sentences, 100,000 sentences, 200,000 sentences, 400,000 sentences, and 800,000 sentences. To examine the relation between update at the SG layer and update at the output layer, at each point in training we computed the update of activation at the output layer (summed over all role-filler pairs) analogously to the SG activation.

To simulate semantic priming effects on N400s during near-chance lexical decisions (simulation 15), we examined the model when it had been presented with just 10,000 sentences, and failed to understand words and sentences, that is, to activate the corresponding units at the output layer (Fig. 5a). At this stage

in training, we simulated semantic priming, associative priming and semantic incongruity in sentences, as described above.

To simulate the interaction between incongruity and repetition (simulation 16), all sentences from the simulation of semantic incongruity (above) were presented twice, in two successive blocks (running through the first presentation of all sentences before running through the second presentation) with connection weights being adapted during the first round of presentations (learning rate = 0.01). 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 second block, and that the incongruent and congruent version of each sentence directly followed each other. The order of conditions (whether the incongruent or congruent version of each sentence was presented first) was counterbalanced across models and items—for half of the models, the incongruent version was presented first for one half of the items; 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[45–47]. Because the SG activation at any given time represents the model's implicit prediction of the semantic features of all aspects of the described event, the change in activation induced by the next word corresponds to the prediction error contained in the previous representation (at least as far as revealed by that next word). Thus, in accordance with the view that prediction errors drive learning, we used a temporal difference 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 activation at the next word serves as the target for the SG activation at the current word, so that the error signal becomes the difference in activation between both words: $SG_n - SG_{n-1}$ (see Supplementary Discussion). There were 10 items per condition; the SU was computed during the first and second presentation of each sentence as the SG difference between the presentation of action (word $n-1$) and object (word n).

Simple recurrent network model. Details on the SRN simulations are described in Supplementary Methods 3.

Statistics. Details on the statistics are reported in Supplementary Methods 4.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary.

Code availability. The computer code used to run the simulations and analyse the results is available on GitHub (https://github.com/milenarabovsky/SG_model).

Data availability. The datasets generated and analysed during the current study are available on GitHub (https://github.com/milenarabovsky/SG_model).

Received: 31 October 2016; Accepted: 19 July 2018;
Published online: 27 August 2018

References

1. Kutas, M. & Hillyard, S. A. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* **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. in *The Psychology of Learning and Motivation–Advances in Research and Theory* Vol. 51, 1–44 (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 Univ. Press, Oxford, 2002).
12. McClelland, J. L., St. John, M. F. & 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. *Vis. Res.* **49**, 1295–1306 (2009).
18. Griffiths, T. L., Stevyers, M. & Tenenbaum, J. B. Topics in semantic representation. *Psychol. Rev.* **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. Preprint at https://arxiv.org/abs/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 (Wiley, New York, NY, 2015).
23. Barsalou, L. W. Grounded cognition. *Annu. Rev. Psychol.* **59**, 617–645 (2008).
24. Pulvermüller, F. Words in the brain's language. *Behav. Brain Sci.* **22**, 253–336 (1999).
25. Kutas, M. & Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature* **307**, 101–103 (1984).
26. Van Petten, C. & Kutas, M. Influences of semantic and syntactic context on open- and closed-class words. *Mem. Cogn.* **19**, 95–112 (1991).
27. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).
28. 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).
29. 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).
30. Hagoort, P., Baggio, G. & Willems, R. M. in *The Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 819–836 (MIT, Cambridge, MA, 2009).
31. Barber, H., Vergara, M. & Carreiras, M. Syllable-frequency effects in visual word recognition: evidence from ERPs. *Neuroreport* **15**, 545–548 (2004).
32. Koivisto, M. & Revonsuo, A. Cognitive representations underlying the N400 priming effect. *Cogn. Brain Res.* **12**, 487–490 (2001).
33. Rugg, M. D. The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology* **22**, 642–647 (1985).
34. 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).
35. Kim, A. & Osterhout, L. The independence of combinatory semantic processing: evidence from event-related potentials. *J. Mem. Lang.* **52**, 205–225 (2005).
36. Brouwer, H., Crocker, M. W., Venhuizen, N. J. & Hoeks, J. C. J. A neurocomputational model of the N400 and the P600 in language processing. *Cogn. Sci.* **41**, 1318–1352 (2017).
37. Van Herten, M., Kolk, H. H. J. & Chwilla, D. J. An ERP study of P600 effects elicited by semantic anomalies. *Cogn. Brain Res.* **22**, 241–255 (2005).
38. 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).
39. Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E. & Kutas, M. Multiple effects of sentential constraint on word processing. *Brain Res.* **1146**, 75–84 (2007).
40. Friedrich, M. & Friederici, A. D. N400-like semantic incongruity effect in 19-month-olds: processing known words in picture contexts. *J. Cogn. Neurosci.* **16**, 1465–1477 (2004).
41. 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).
42. Kutas, M. & Iragui, V. The N400 in a semantic categorization task across 6 decades. *Electroencephalogr. Clin. Neurophysiol.* **108**, 456–471 (1998).
43. Gotts, S. J. Incremental learning of perceptual and conceptual representations and the puzzle of neural repetition suppression. *Psychon. Bull. Rev.* **23**, 1055–1071 (2016).
44. McLaughlin, J., Osterhout, L. & Kim, A. Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nat. Neurosci.* **7**, 703–704 (2004).
45. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
46. Friston, K. A theory of cortical responses. *Phil. Trans. R. Soc. Lond.* **360**, 815–836 (2005).
47. McClelland, J. L. in *International Perspectives on Psychological Science* (eds Bertelson, P, Eelen, P. & d'Ydewalle, G.) Vol. 1, 57–88 (Lawrence Erlbaum Associates, Hillsdale, 1994).

48. Besson, M., Kutas, M. & Van Petten, C. An Event-Related Potential (ERP) analysis of semantic congruity and repetition effects in sentences. *J. Cogn. Neurosci.* **4**, 132–149 (1992).
49. 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).
50. Kuperberg, G. R. Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Lang. Cogn. Neurosci.* **31**, 602–616 (2016).
51. 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).
52. Rumelhart, D. E. in *Metaphor and Thought* (ed. Ortony, A.) 71–82 (Cambridge Univ. Press, Cambridge, UK, 1979).
53. 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).
54. 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).
55. Sanford, A. J. & Sturt, P. Depth of processing in language comprehension: not noticing the evidence. *Trends Cogn. Sci.* **6**, 382–386 (2002).
56. Ferreira, F., Bailey, K. G. D. & Ferraro, V. Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* **11**, 11–15 (2002).
57. Dronkers, N. F. et al. Lesion analysis of the brain areas involved in language comprehension. *Cognition* **92**, 145–177 (2004).
58. 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).
59. Bookheimer, S. Functional MRI of language: new approaches to understanding the cortical organization of semantic processing. *Annu. Rev. Neurosci.* **25**, 151–188 (2002).
60. Friederici, A. D. Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* **6**, 78–84 (2002).
61. 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. USA* **94**, 14792–14797 (1997).
62. 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).
63. 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).
64. 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).
65. 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).
66. 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).
67. Osterhout, L. & Holcomb, P. J. Event-related brain potentials elicited by syntactic anomaly. *J. Mem. Lang.* **31**, 785–806 (1992).
68. Regel, S., Gunter, T. C. & Friederici, A. D. Isn't it ironic? An electrophysiological exploration of figurative language processing. *J. Cogn. Neurosci.* **23**, 277–293 (2010).
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. Nieuwland, M. S. & Kuperberg, G. R. When the truth is not too hard to handle. *Psychol. Sci.* **19**, 1213–1218 (2008).
76. Staab, J., Urbach, T. & Kutas, M. Negation processing in context is not (always) delayed. *Cent. Res. Lang. Tech. Rep.* **20**, 3–34 (2009).
77. van Berkum, J. J., Hagoort, P. & Brown, C. M. Semantic integration in sentences and discourse: evidence from the N400. *J. Cogn. Neurosci.* **11**, 657–671 (1999).
78. Nieuwland, M. S. & Van Berkum, J. Ja When peanuts fall in love: N400 evidence for the power of discourse. *J. Cogn. Neurosci.* **18**, 1098–1111 (2006).
79. 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).
80. Rumelhart, D. E. & Todd, P. M. in *Attention and Performance XIV* 3–30 (MIT, Cambridge, MA, 1993).
81. McClelland, J. L. & Rogers, T. T. The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.* **4**, 310–322 (2003).
82. Pennington, J., Socher, R. & Manning, C. in *Proc. 2014 Conf. Empiric. Methods Natur. Lang. Process. (EMNLP)* 1532–1543 (Association for Computational Linguistics, 2014).
83. Altmann, G. T. M. & Kamide, Y. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition* **73**, 247–264 (1999).
84. 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).

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 M.R. We thank R. Levy, S. Frank and the members of the PDP lab at Stanford University for helpful discussion. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0406-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.R. or J.L.M.

nature research

Corresponding author(s): Milena Rabovsky, James L. McClelland

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	MATLAB_R2015a MATLAB toolbox PDPTool Version 3 (http://web.stanford.edu/group/pdplab/pdphandbookV3/handbookap1.html#x25-143000A) The code used to run the simulations is available on https://github.com/milenarabovsky/SG_model
Data analysis	MATLAB_R2015a, R version 3.3.3 (2017-03-06) The code used to analyze the results is available on https://github.com/milenarabovsky/SG_model

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated and analyzed during the study are available on https://github.com/milenarabovsky/SG_model

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	There is much less noise in the simulations as compared to empirical experiments such that the small sample size (10 runs of the model and 8 to 16 items per condition) is adequate.
Data exclusions	No data were excluded from the analysis.
Replication	There was no replication. However, we verified the reproducibility of the findings by testing each effect in 10 independently initialized and trained models.
Randomization	The only situation where randomization of presentation order plays a role is in the blockwise repetition simulation. The entire set of congruent and incongruent sentences was presented a first time before all sentences were presented again a second time. The order of presentation within each block was randomized independently for each model with the restriction that the order of presentation of the set of sentences was the same for the first and the second block, and that the incongruent and congruent version of a sentence directly followed each other, with the order of conditions counterbalanced across models and items.
Blinding	Blinding was not relevant to our study, because no behavioral or biological data were collected or analysed.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging