nature human behaviour

SUPPLEMENTARY INFORMATION

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

In the format provided by the authors and unedited.

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

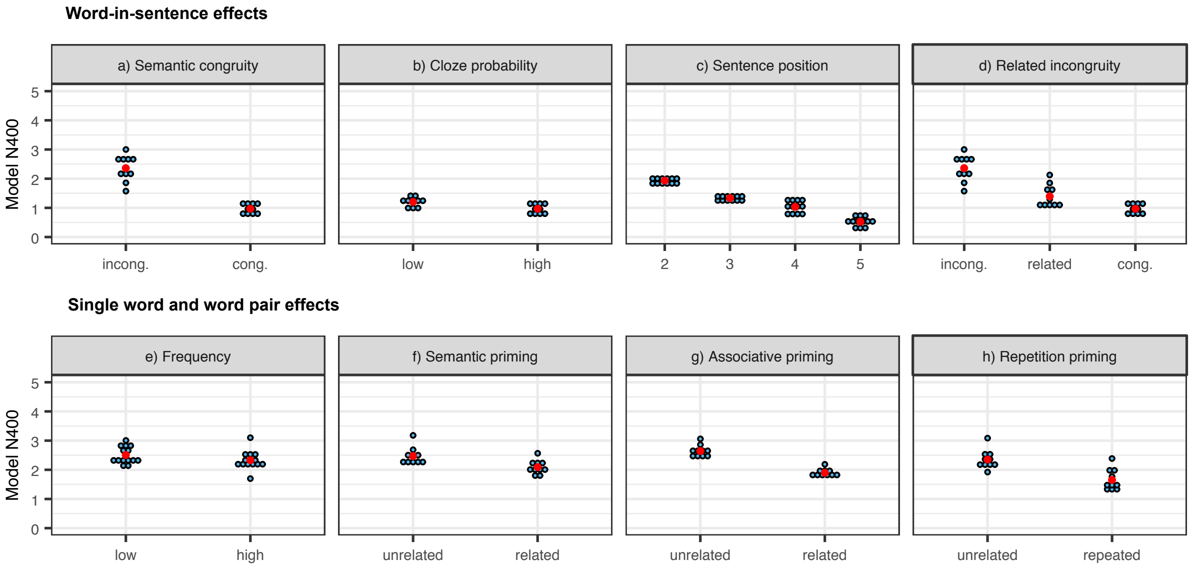
Milena Rabovsky*, **Steven S. Hansen** and **James L. McClelland***

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

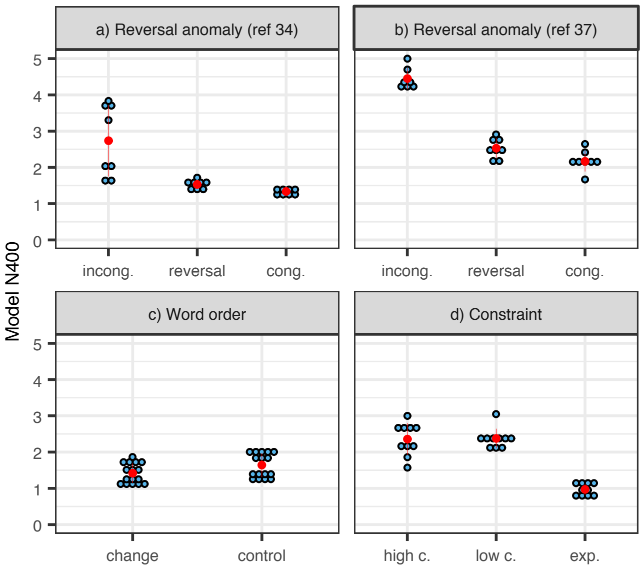
Supplementary materials

Table of content

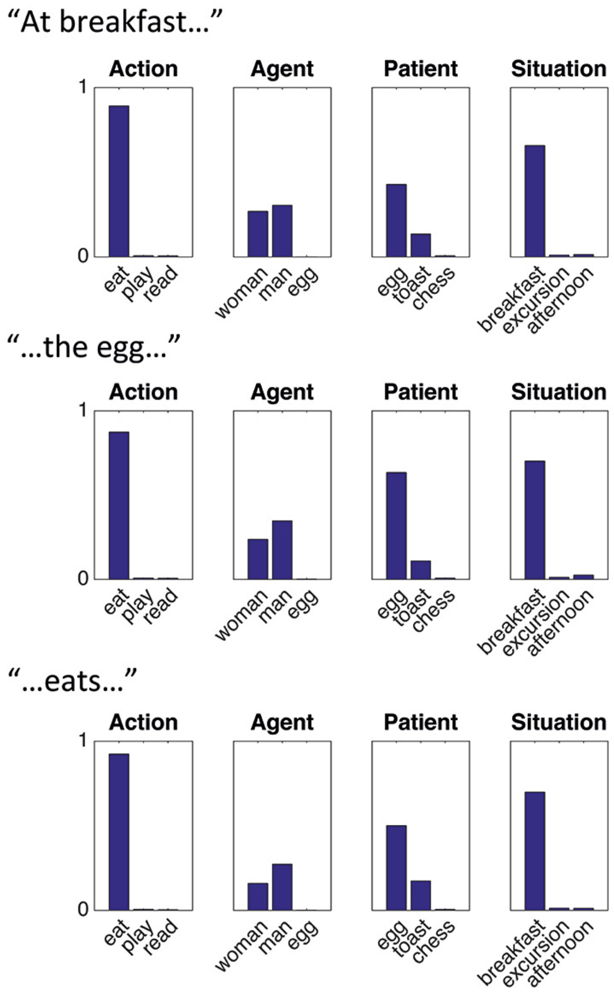
Supplementary Figures



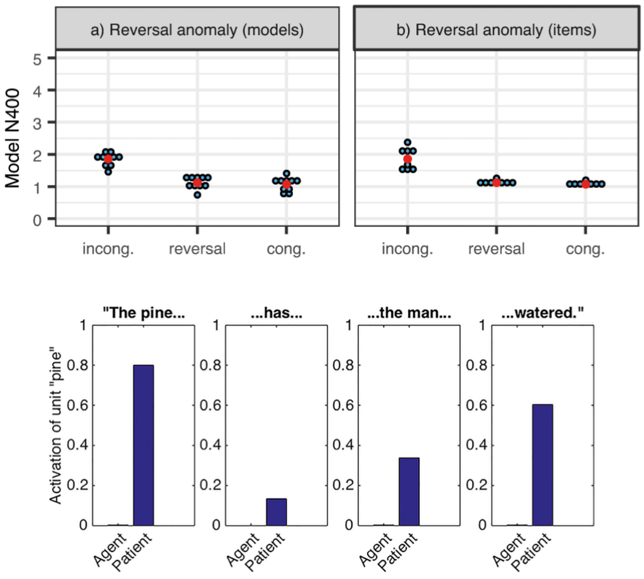
Supplementary Figure 1. Simulation results for the basic effects (by item). Displayed is the model's N400 correlate, i.e. the update of the Sentence Gestalt layer activation – the model's probabilistic representation of sentence meaning - induced by the new incoming word. Cong., congruent; incong., incongruent. Each blue dot represents the results for one item, averaged across 10 independent runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM (sometimes invisible because bars may not exceed the area of the red dot). Item-based analyses: **a**, semantic incongruity (*n* = 10 items/ condition): $t_2(9) = 11.24$, $p < .001$, $d = 3.55$, 95% CI [1.11, 1.67]; **b**, cloze probability (*n* = 10 items): $t_2(9) = 6.42$, $p < .001$, $d = 2.03$, CI 95% [.15, .32]; **c**, position in sentence (*n* = 12 items): $t_2(11) = 43.54$, $p <.001$, $d = 12.57$, 95% CI [.57, .63] from second to third sentence position; $t_2(11) = 4.66$, $p =.0018$, $d = 1.34$, 95% CI [.16, .44] from third to fourth position; $t_2(11) = 12.65$, $p <.001$, $d = 3.65$, 95% CI [.42, .60] from fourth to fifth position; **d**, categorically related incongruities (*n* = 10 items) were larger than congruent, $t_2(9) = 3.31$, $p = .018$, $d = 1.05$, 95% CI [.13, .71], and smaller than incongruent continuations, $t_2(9) = 12.44$, $p < .001$, $d = 3.94$, 95% CI [.79, 1.14]; **e**, lexical frequency (*n* = 14 items): $t_2(13) = 3.26$, $p = .0062$, $d = .87$, 95% CI [.06, .30]; **f**, semantic priming (*n* = 10 items): $t_2(9) = 8.92$, $p < .001$, $d = 2.82$, 95% CI [.28, .48]; **g**, associative priming (*n* = 10 items): $t_2(9) = 18.42$, $p < .001$, $d = 5.82$, 95% CI [.65, .84]; **h**, immediate repetition priming (*n* = 10 items): $t_2(9) = 18.93$, $p < .001$, $d = 5.99$, 95% CI [.62, .79].



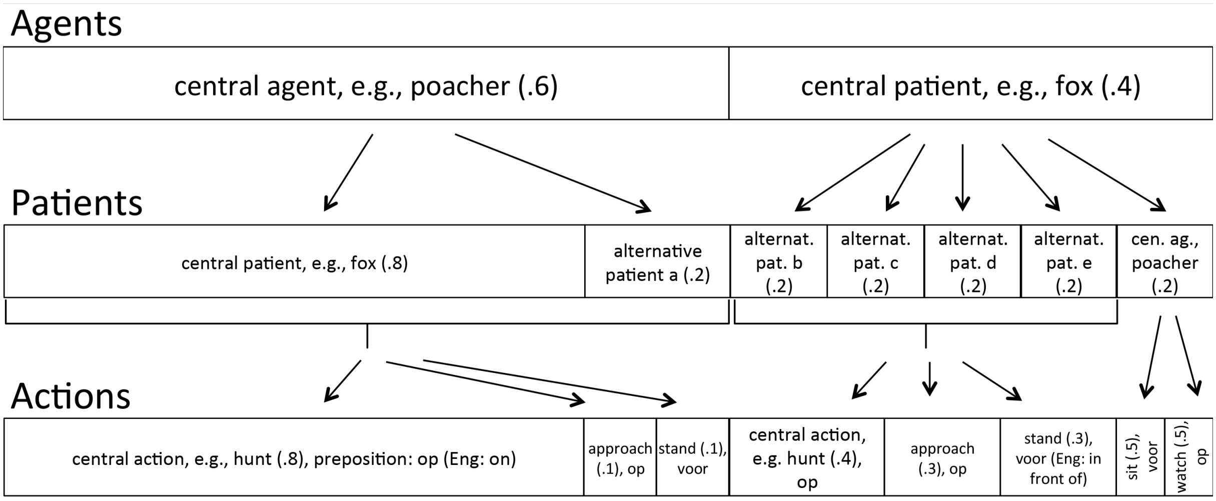
Supplementary Figure 2. Simulations results concerning the specificity of the N400 effect (by item). Each blue dot represents the results for one item, averaged across 10 independent runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM. Incong, incongruent; reversal, reversal anomaly; cong., congruent; change, changed word order; control, normal word order; high c., unexpected high constraint; low c., unexpected low c.; exp., expected. Item-based analyses: **a**, reversal anomaly in the standard model[1] (*n* = 8 items): $t_2(7) = 5.67$, $p = .0023$, $d = 2.0$, 95% CI [.12, .28], for comparison between congruent condition and reversal anomaly; $t_2(7) = 3.56$, $p = .028$, $d = 1.26$, 95% CI [.40, 2.0] for comparison between reversal anomaly and incongruent, and $t_2(7) = 4.21$, $p = .012$, $d = 1.49$, 95% CI [.61, 2.19] for comparison between congruent and incongruent. **b**, reversal anomaly where both participants can be agents[2] (*n* = 8 items). These results are from a model trained on a different environment (see main text), 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 it is considerably larger in the incongruent condition. Congruent versus reversal anomaly: $t_2(7) = 2.81$, $p = .052$, $d = 0.99$, 95% CI [.06, .67]; congruent versus incongruent: $t_2(7) = 22.37$, $p < .001$, $d = 7.91$, 95% CI [2.10, 2.60]; reversal anomaly versus incongruent: $t_2(7) = 17.92$, $p < .001$, $d = 6.34$, 95% CI [1.73, 2.25]. **c**, change in word order (*n* = 10 items). SU was slightly larger for normal versus changed order; significant over models (see caption of Fig. 3 in main text), but not items, $t_2(9) = 1.56$, $p = .14$, $d = .39$, 95% CI [-.08, .53]. **d**, constraint for unexpected endings (*n* = 10 items). SU did not differ between unexpected high constraint and unexpected low constraint constraint, $t_2(9) = 0.12$, $p = .91$, $d = .04$, 95% CI [-.27, .30]. For expected endings it was lower than for both, unexpected high constraint, $t_2(9) = 11.24$, $p < .001$, $d = 3.55$, 95% CI [1.11, 1.67], and unexpected low constraint, $t_2(9) = 23.33$, $p < .001$, $d = 7.38$, 95% CI [1.27, 1.54].



Supplementary Figure 3. Processing reversal anomalies. Activation of selected output units while the model processes a sentence from the first reversal anomaly simulation[1] (Simulation 9): “At breakfast, the egg eats…”. Note that the model continues to represent the egg as the patient (not the agent) of eating, even after the word “eat” has been presented.

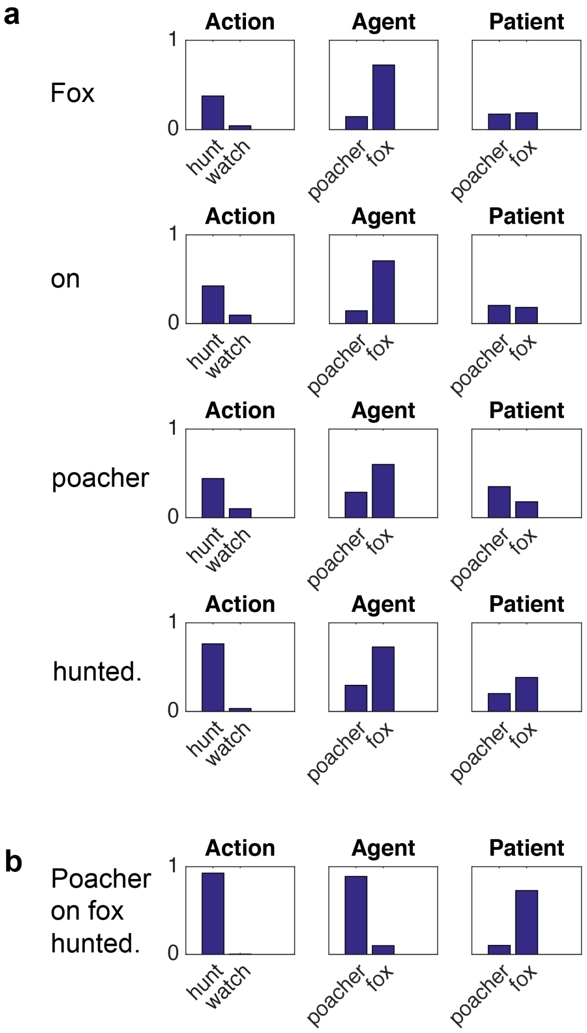


Supplementary Figure 4. Simulation results for a type of reversal anomaly where the relationship between two noun phrases is established prior to encountering the verb (e.g., lit: "The javelin has the athletes thrown"[3]; see Supplementary Methods 1 for details); the simulation was conducted with a model trained with Dutch word order. Cong., congruent; incong., incongruent; reversal, reversal anomaly. **a**, Each blue dot represents the results for one independent run of the model, averaged across the eight items per condition. **b,** Each blue dot represents the results for one item, averaged across 10 independent runs of the model. The red dots represent the means for each condition, and red error bars represent +/- SEM. Results are similar as for the other reversal anomaly simulation: $t_1(9) = 1.69$, $p = .38$, $d = .53$, 95% CI [-.02, .12], $t_2(7) = 12.67$, $p < .001$, $d = 4.48$, 95% CI [.04, .06], for the comparison between congruent condition and reversal anomaly; $t_1(9) = 13.31$, $p < .001$, $d = 4.21$, 95% CI [.61, .86], $t_2(7) = 6.76$, $p < .001$, $d = 2.39$, 95% CI [.48, 1.0], for the comparison between reversal anomaly and incongruent condition, and $t_1(9) =12.18$, $p < .001$, $d = 3.85$, 95% CI [.65, .94], $t_2(7) = 7.36$, $p < .001$, $d = 2.60$, 95% CI [.54, 1.05], for the comparison between congruent and incongruent condition. **Bottom.** Activation of the unit "pine" in response to the Agent and Patient probe while the model processes a sentence from this reversal anomaly simulation, literally "The pine has the man watered." (i.e., "The pine has watered the man." with Dutch word order). As for the "eggs" type anomaly sentences, the model represents the pine as the patient instead of the agent of the event throughout the sentence.

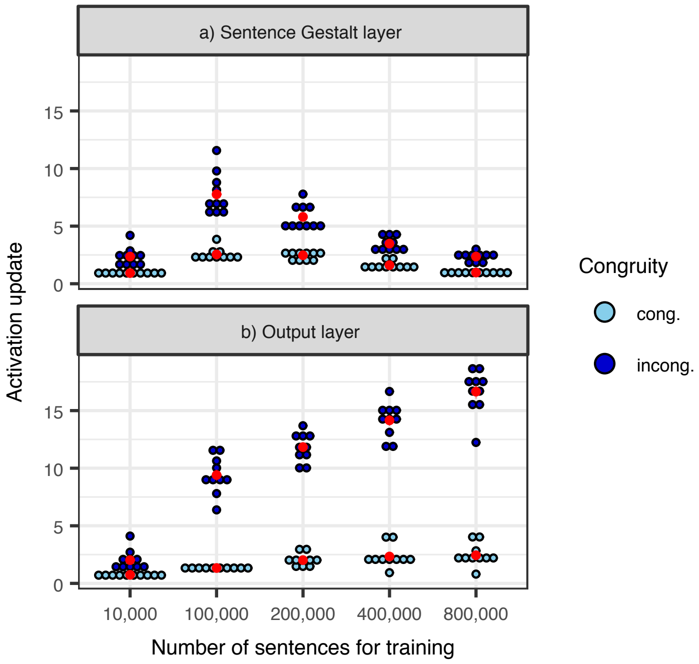

Agents
central agent, e.g., poacher (.6)
central patient, e.g., fox (.4)
Patients
central patient, e.g., fox (.8)
alternative patient a (.2)
alternat. pat. b (.2)
alternat. pat. c (.2)
alternat. pat. d (.2)
alternat. pat. e (.2)
cen. ag., poacher (.2)
Actions
central action, e.g., hunt (.8), preposition: op (Eng: on)
approach (.1), op
stand (.1), voor
central action, e.g. hunt (.4), op
approach (.3), op
stand (.3), voor (Eng: in front of)
sit (.5), voor
watch (.5), op

Sentence structure: 70% active: [Agent] [Preposition] [Patient] [Action]
30% passive: [Patient] [was by] [Agent] [Action]

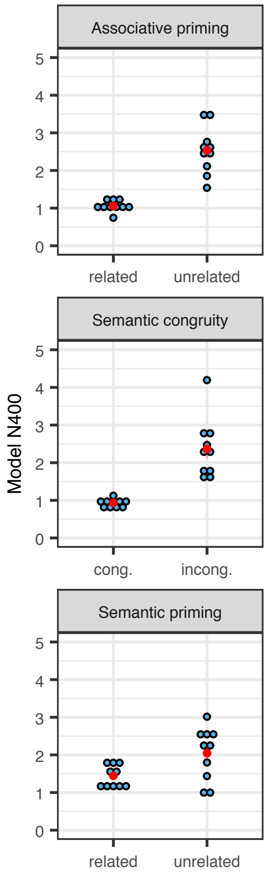
Supplementary Figure 5. Example scenario from the training environment designed to simulate N400 amplitudes during reversal anomalies with two animate event participants such as “The fox on the poacher hunted.”[2] Eight analogous scenarios were added to the training environment (see Methods and Supplementary Fig. 12; with Dutch word order) to be able to assess the reliability of the effects across items. English words are used as labels for the scenario participants to help the reader align the design of the scenarios with a natural sentence example; the central agent, patient, and action are the most frequent filler of each role, here labeled ‘poacher’, ‘fox’ and ‘hunt’ respectively. The alternative actions (approach, stand, sit, and watch) were shared across all eight scenarios, with counterbalanced assignment to the different slots in the scheme above (e.g., in another scenario, ‘watch’ and ‘sit’ switch positions with ‘approach’ and ‘stand’). The alternative actions are intended to capture the existence of unspecific actions that can be performed by a wide variety of agents towards a wide variety of patients. To capture the impression that many of the central agents in the experimental sentences were relatively unlikely to be patients in events involving the respective central patient, but could very well be patients in a variety of alternative events, the central agents filled the patient role in different scenarios. Specifically, in two groups of four scenarios each, for each scenario, the central agents from the remaining three scenarios within the group, and additionally one of the central agents from a scenario from the other group, were used as alternative patients b, c, d, and e (see scheme above). Furthermore, all the central agents also occurred as patients in events involving the generic actions ‘like’ and ‘look at’ from the pre-existing part of the environment (see Supplementary Fig. 12). Similarly, while the central patients were slightly less likely to occur as agents than the central agents within the scenarios, they could additionally occur as agents in events involving the generic action ‘like’. This is intended to capture the impression that on average the central patients in the experimental sentences were fairly likely to be agents (with some e.g. ‘the fox’ more likely than others, e.g. ‘the patient’). When the central agent was the agent of the event, the alternative patient (alternative patient a in the scheme above) was shared across two scenarios (i.e., there were overall four such patients). Each concept is represented by four semantic features at the output layer. For current purposes, the crucial point in assigning the semantic features was to avoid any systematic differences between the central agents and the central patients. The central agent and the central patient (‘fox’ and ‘poacher’, above) each have three unique features and one feature labeled ‘active’ which is shared with the agents in the pre-existing part of the training environment (see Supplementary Table 1). All actions have three unique features and one feature labeled ‘action’, which is shared with the actions in the pre-existing part of the training environment. The prepositions are not associated with any semantic features. There were two different propositions (these can be thought of as corresponding to the Dutch ‘op’ and ‘voor’, which were used frequently in the experimental materials). Both occurred with half of the central actions and overall occurred equally frequently. The resulting model has 71 units at the input layer and 182 units at the probe and output layer.



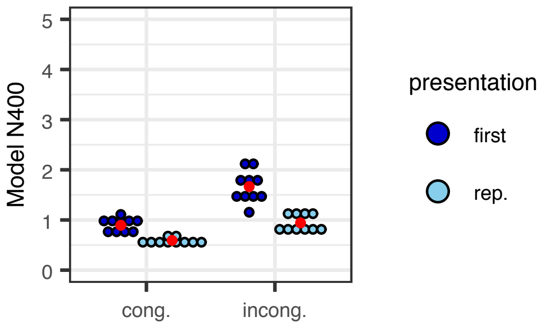
Supplementary Figure 6. The model's interpretation of sentences from the simulation of reversal anomalies with two animate event participants[2] (see Supplementary Methods 2 and Supplementary Fig. 5 for details). The simulation was conducted with a model trained with Dutch word order and the example sentences shown are literal translations from Dutch. For the visualization of the model's interpretation, English words are used to help the reader map the displayed activations to natural sentence examples. However, it is important to note that the activations are averaged across the respective event participants ('fox' representing the central patient, 'poacher' representing the central agent, and 'hunt' representing the central action) in eight analogous scenarios. **a**, The model's average activation of units representing the central agent ('poacher' in the example), the central patient ('fox' in the example), the central action ('hunt' in the example) as well as an alternative action ('watch' in the example) when probed for the Action, Agent and Patient role over the course of a reversal anomaly sentence describing an event where the central patient does the central action to the central agent ('The fox on the poacher hunted.'). **b**, The model's average activation of the same units after the presentation of the verb in a congruent sentence describing an event where the central agent does the central action to the central patient ('The poacher on the fox hunted.'). Please note that the model's representation differs between the congruent and the reversal anomaly sentence. While the interpretation in the congruent sentence is unambiguous and clear, the representation in the reversal anomaly sentence reflects a state of unresolved conflict between different cues, demonstrating the model's joint sensitivity to event probability and word order constraints.



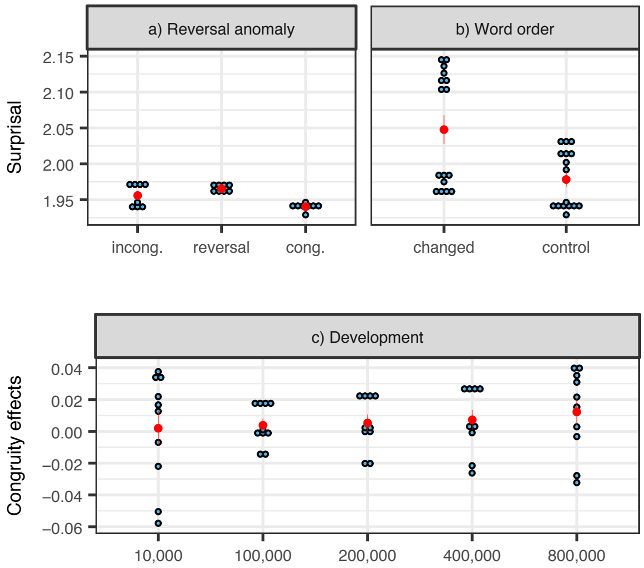
Supplementary Figure 7. Development across training (by item). Semantic incongruity effects as a function of the number of sentences the model has been exposed to. Each light blue or dark blue dot represents the results for one item (10 per condition), averaged across 10 independent runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM. **a,** Semantic update at the model's hidden Sentence Gestalt layer shows at first an increase and later a decrease with additional training, in line with the developmental trajectory of the N400. Item-bases analyses: The size of the effect (i.e. the numerical difference between the congruent and incongruent condition) differed between all subsequent time points: $t_2(9) = 6.94$, $p < .001$, $d = 2.19$, 95% CI [2.55, 5.02] between 10,000 and 100,000 sentences; $t_2(9) = 10.05$, $p < .001$, $d = 3.18$, 95% CI [1.45, 2.29] between 100,000 and 200,000 sentences; $t_2(9) = 6.87$, $p < .001$, $d = 2.17$, 95% CI [.98, 1.95] between 200,000 and 400,000 sentences; $t_2(9) = 3.70$, $p = .02$, $d = 1.17$, 95% CI [.19, .78] 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.



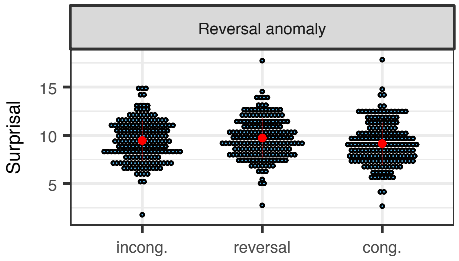
Supplementary Figure 8. Semantic update effects at a very early stage in training (by item). Cong., congruent; incong., incongruent. Even at a low level of performance (see Fig. 5a in the main text for illustration), there are robust effects of associative priming (top), semantic congruity in sentences (middle), and semantic priming (bottom). Each blue dot represents the results for one item, averaged across 10 independent runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM. Item-based analyses: associative priming (top): $t_2(9) = 7.31$, $p < .001$, $d = 2.31$, 95% CI [1.02, 1.94]; semantic congruity in sentences (middle): $t_2(9) = 5.74$, $p < .001$, $d = 1.81$, 95% CI [.86, 1.99]; semantic priming (bottom): $t_2(9) = 3.79$, $p = .0043$, $d = 1.20$, 95% CI [.24, .96].



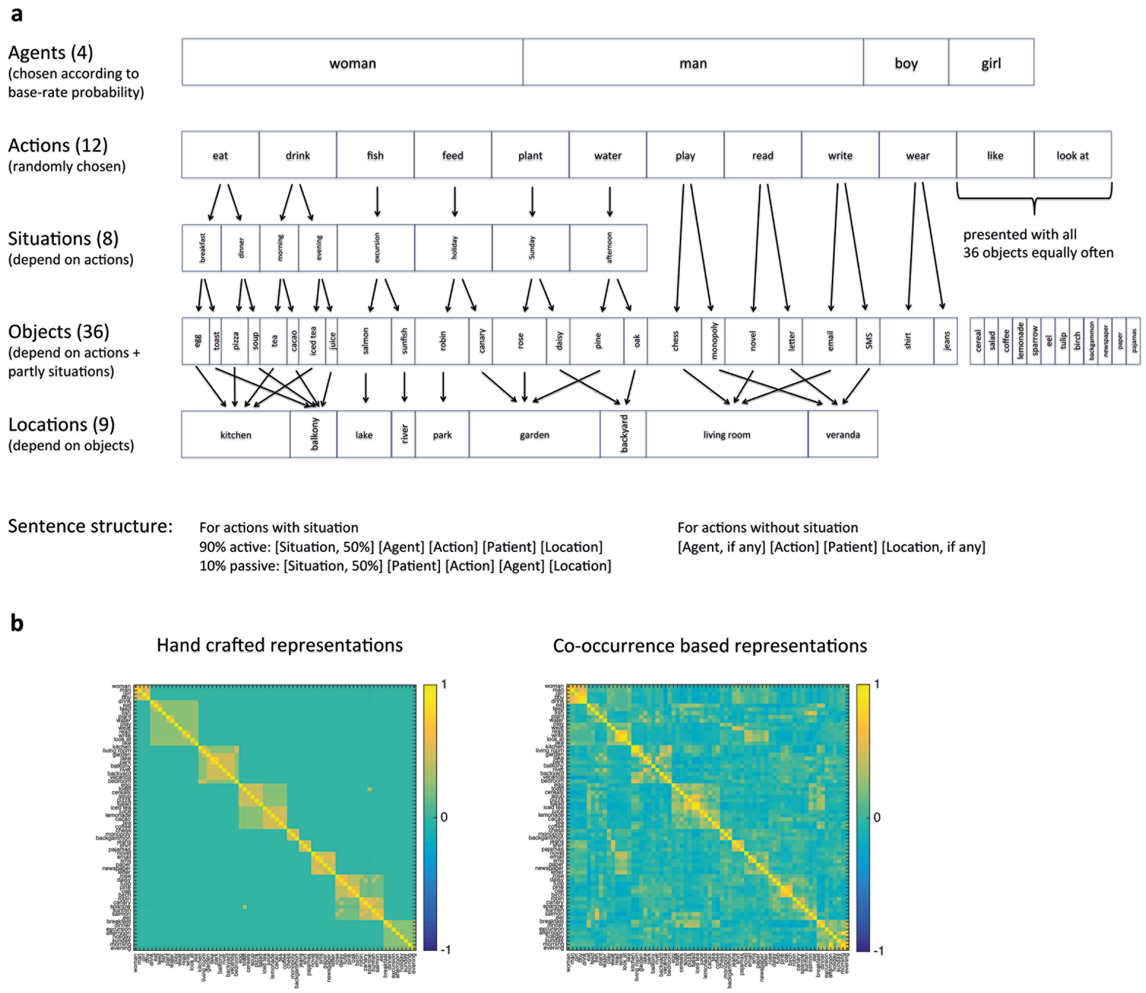
Supplementary Figure 9. Simulation of the interaction between delayed repetition and semantic incongruity (by item). Each dark blue or light blue dot represents the results for one item, averaged across 10 runs of the model; red dots represent means for each condition, and red error bars represent +/- SEM. Item-based analyses: There were significant main effects of congruity, $F_2(1,9) = 115.66$, $p < .001$, $\eta_p^2 = .928$ and repetition, $F_2(1,9) = 109.78$, $p < .001$, $\eta_p^2 = .924$, and a significant interaction between both factors, $F_2(1,9) = 120.86$, $p < .001$, $\eta_p^2 = .931$. Post-hoc comparisons showed that even though the repetition effect was larger for incongruent as compared to congruent sentence completions (i.e., incongruent [first – repetition] > congruent [first – repetition], $t_2(9) = 10.99$, $p < .001$, $d = 3.48$, 95% CI [.34, .51], it was significant in both conditions, $t_2(9) = 6.90$, $p < .001$, $d = 2.18$, 95% CI [.20, .40], for the congruent completions, and $t_2(9) = 12.02$, $p < .001$, $d = 3.80$, 95% CI [.59, .86] for the incongruent completions.



Supplementary Figure 10. Simulation results from a simple recurrent network model (SRN) trained to predict the next word based on the preceding context (by item). Each blue dot represents the results for one item, averaged across 10 runs of the model; red dots represent means for each condition, and red error bars represent +/- SEM. Item-based analyses: **a**, reversal anomaly[1]: $t_2(7) = 7.83$, $p < .001$, $d = 2.77$, 95% CI [.018, .033] for the comparison between congruent and reversal anomaly; $t_2(7) = 2.98$, $p = .062$, $d = 1.05$, 95% CI [.003, .028] for the comparison between congruent and incongruent; $t_2(9) = 1.57$, $p = .48$, $d = .55$, 95% CI [-.005, .024] for the comparison between incongruent and reversal anomaly. **b**, word order: $t_2(15) = 6.73$, $p < .001$, $d = 1.68$, 95% CI [.048, .092]. **c**, congruity effect on surprisal as a function of the number of sentences the model has been exposed to: $t_2(9) = .15$, $p = 1.0$, $d = .048$, 95% CI [-.027, .031] for the comparison between 10,000 and 100,000 sentences; $t_2(9) = 1.08$, $p = 1.0$, $d = .34$, 95% CI [-.0015, .0043] for the comparison between 100,000 and 200,000 sentences; $t_2(9) = 1.78$, $p = .44$, $d = .56$, 95% CI [-.0005, .0045] for the comparison between 200,000 and 400,000 sentences; $t_2(9) = 1.93$, $p = .36$, $d = .61$, 95% CI [-.0008, .011] for the comparison between 400,000 and 800,000 sentences.



Supplementary Figure 11. Simulation results from a simple recurrent network (SRN) implementation by T. Mikolov[4] trained by S. Frank on 23M sentences from a web corpus. Incong., incongruent; cong., congruent; reversal, reversal anomaly. The simulation experiment consisted in the presentation of materials from the reversal anomaly experiment by Kuperberg and colleagues[1] which we requested from the authors (there are a few slight differences in the materials due to an issue with retrieving the original stimuli, but the materials largely overlap and resulted in the same pattern of results; G. Kuperberg, personal communication). Each black dot represents the results for one item, averaged over three runs of the model; the red dots represent the means for each condition, and red error bars represent +/- SEM. Results resemble those from the SRN that we trained on the same corpus as the SG model (Fig. 7 and Supplementary Fig. 10) in that word surprisal was large in the reversal anomaly condition, numerically even larger than in the incongruent condition. There were 3 runs of the model and 180 items in each condition (1 less in the incongruent condition because the model did not know one of the words in this condition, "curtseys") so that we report statistical results from the item analyses: $t_2(179) = 11.76$, $p < .001$, $d = .88$, 95% CI [.48, .67] for the comparison between congruent condition and reversal anomaly; $t_2(178) = 1.29$, $p = .59$, $d = .10$, 95% CI [-.13, .66] for the comparison between incongruent condition and reversal anomaly, and $t_2(178) =1.45$, $p= .45$, $d = .11$, 95% CI [-.11, .73] for the comparison between congruent and incongruent condition. We thank Stefan Frank for performing the simulation and sharing the results with us!



Supplementary Figure 12. a, The standard sentence/ event generator used to train the model. Bar width corresponds to relative probability. First, one out of twelve actions is chosen with equal probability. Then, for every action except one ("look at") an agent is chosen ("woman" and "man" each with a probability of .4, "boy" and "girl" with a probability of .1). Next, a situation is chosen depending on the action. Some actions can occur in two possible situations, some in one, and some without a specified situation. Even if an action occurs in a specific situation, the corresponding word is presented only with a probability of .5 in the sentence while the situation is always part of the event representation. Then, depending on the action (and in the case that an action can occur in two possible situations, depending on the situation) an object/patient is chosen. For each action or situation (except for "like" and "look at" for which all 36 objects are chosen equally often) there is a high probability and a low probability object (if the agent is "man" or "woman", the respective high/low probabilities are .7/.3, if the agent is "girl" or "boy", the probabilities are .6/.4). The high and low probability objects occurring in the same specific action context are always from the same semantic category, and for each category, there is a third object which is never presented in that action context and instead only occurs in the unspecific "like" or "look at" contexts (to enable the simulation of categorically related incongruities; these are the twelve rightmost objects in the figure; here bar width is larger than probability to maintain readability). Possible sentence structures are displayed below. **b**, Similarity matrices of the hand-crafted semantic representations used for the current model (left) and representations based on a principal component analysis on word vectors derived from co-occurrences in large text corpora (GloVe^5). The correlation between the matrices is r = .73.

Supplementary Table

Supplementary Table 1

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

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

Toast	consumable, food, brown, toast
Cereals	consumable, food, healthy, cereals
Soup	consumable, food, in bowl, soup
Pizza	consumable, food, round, pizza
Salad	consumable, food, light, salad

Iced tea	consumable, drink, from leaves, iced tea
Juice	consumable, drink, from fruit, juice
Lemonade	consumable, drink, sweet, lemonade
Cacao	consumable, drink, with chocolate, cacao
Tea	consumable, drink, hot, tea
Coffee	consumable, drink, activating, coffee

Chess	game, entertaining, strategic, chess
Monopoly	game, entertaining, with dice, monopoly
Backgammon	game, entertaining, old, backgammon

Jeans	garment, to cover body, for legs, jeans
Shirt	garment, to cover body, for upper part, shirt
Pajamas	garment, to cover body, for night, pajamas

Novel	contains language, contains letters, art, novel
Email	contains language, contains letters, communication, email
SMS	contains language, contains letters, communication, short, SMS
Letter	contains language, contains letters, communication, on paper, letter
Paper	contains language, contains letters, scientific, paper
Newspaper	contains language, contains letters, information, newspaper

Rose	can grow, has roots, has petals, red, rose
Daisy	can grow, has roots, has petals, yellow, daisy
Tulip	can grow, has roots, has petals, colorful, tulip

Pine	can grow, has roots, has bark, green, pine
Oak	can grow, has roots, has bark, tall, oak
Birch	can grow, has roots, has bark, white bark, birch

Robin	can grow, can move, can fly, red, robin
Canary	can grow, can move, can fly, yellow, canary
Sparrow	can grow, can move, can fly, brown, sparrow

Sunfish	can grow, can move, can swim, yellow, sunfish
Salmon	can grow, can move, can swim, red, salmon
Eel	can grow, can move, can swim, long, eel

By	passive voice (activated together with the deep subject, e.g., ‘by the man’)
Was	passive voice (activated together with the verb, e.g., ‘was played’)
During/at	no output units (activated together with situation words, e.g., ‘at breakfast’)
In	no output units (activated together with location words, e.g., ‘in the park’)

Supplementary Notes

Supplementary Note 1

In our model's N400 correlate, the effect of congruity is much larger than the effect of cloze probability. This result seems to contrast with studies directly comparing low cloze plausible and low cloze implausible continuations and reporting that the N400 is mainly influenced by cloze probability and much less by plausibility[6]. However, note that in the model's environment, the probability in the low cloze condition was comparatively high (.3), considerably higher than the near-to-zero cloze probability in the mentioned studies[6], and that low cloze and high cloze continuations in the model environment share semantic features, contributing to the comparatively small effect of cloze probability.

Supplementary Note 2

Consistent with the original SG simulations[7], the model's interpretations are sensitive both to event probability constraints and word-order constraints. (As discussed in the main text, we use the phrase 'event probability constraints' 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 occurrence of 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 by this usage, the event probability constraints on the fillers of the Agent, Patient, and Action roles would be the same as well.) In reversal anomaly sentences, word order and event probabilities conflict. The model's representation when processing reversal anomalies such as "The fox on the poacher hunted" (Simulation 11) correspondingly reflects uncertainty and ends up in an inconclusive state (see Suppl. Fig. 6a).

The model starts by favoring the interpretation of the fox as the agent of the event and is uncertain about the patient. When "the poacher" is presented, the model slightly prefers the poacher as the patient (which makes sense based on syntax, i.e. word order and the preposition "on" without passive marker) but at the same time keeps "hunted" as the most probable action (which makes sense based on event probability) and also still maintains the possibility that the poacher may be the agent and the fox may be the patient (counter to the syntactic cues but consistent with event probability). When the final word "hunted" is presented, the model continues to exhibit cue conflict. There is a shift in the probabilities for the filler of the patient role from a tendency towards the poacher to a slight preference for the fox, which makes sense based on event probabilities (the fox is always the patient in events additionally involving a poacher and hunting). At the same time the model maintains a high probability for the fox being the agent based on syntactic cues (word order and active voice) and also maintains the possibility that the poacher could be either the agent (in line with event probabilities) or the patient (in line with the syntax). In short, when the constraints based on word order and event probabilities agree, as in the control sentences, there is a strong preference for one specific interpretation, with very high activations for the correct role fillers and very low activations for alternative role filler pairs (Suppl. Fig 6b). When the constraints conflict, the model may remain in a state of relative uncertainty and indetermination, in line with the notion that representations during human language comprehension can remain underspecified[8] (Suppl. Fig. 6a).

We also examined the model's capacity to assign roles correctly when the reversal anomaly context (e.g., 'the fox on the poacher') was followed by a verb that it had experienced in such contexts during training (e.g. 'watched'; see Supplementary Fig. 5 for details on the training environment). The model performed correctly across all of the scenarios tested, in that the correct filler was most active in each of the Agent, Patient, and Action roles. The model does not strongly pre-activate 'watch' upon presentation of 'the fox on the poacher…' (which would result in a larger SU upon presentation of 'hunted') due to the event probability constraints, which favor hunting in events involving poachers and foxes.

Supplementary Note 3

There is considerable evidence that representations formed during language comprehension are sometimes influenced by event probabilities[9]. To directly investigate the claim that the N400 corresponds to the formation of such representations one could combine N400 measurements with comprehension questions, probing, for example, the comprehension of role-reversed sentences which have been shown to lead to a high rate of role assignment errors (e.g., "The dog was bitten by the *man*.")[9]. This would allow for direct examination of the co-variation between N400s and event probability based comprehension.

Specifically, when presented with sentences such as "The dog was bitten by the man.", in instances where participants understand the dog to be the agent and the man to be the patient, N400 amplitudes on *man* should be small. For interpretations in which the understanding is in line with syntactic conventions, the situation is more complicated. Depending on participant's prior experience and details of the relevant word order and event probability information, there might be instances where the syntactically-specified assignment was understood immediately and instances where participants temporarily process the sentence in line with event probability constraints or experience uncertainty that gets resolved later in the process. Thus, in these cases, one might expect to either see a large N400 (reflecting immediate interpretation in line with syntactic conventions and thus considerable semantic update) or a small N400 (reflecting a temporary event probability based interpretation or uncertainty) accompanied by an indication of an enhanced controlled update process later in the process (which might be reflected in increased P600 amplitudes).

Supplementary Note 4

Negation is typically used to deny a supposition, and in the absence of discourse context, this supposition must be grounded in general knowledge[10]. Thus, when used in short and isolated sentences, negation is typically used to deny something that is part of an invoked schema (e.g., "a whale is not a fish"). "Robin" does not invoke a schema which includes semantic features of "vehicle" so that "A robin is not a vehicle" is not an expected sentence meaning, even though it is true. On the other hand, "robin" does invoke a schema which includes semantic features of "bird" so that something that is part of the schema of "bird" might be expected to be denied (e.g., "A robin is not a bird that flies south during winter" is fine).

Supplementary Note 5

Aspects of the model environment relevant to the simulation of the influence of a word's position in the sentence (Simulation 3). Given a specific situation, the conditional probability of the presented agent (“man”; at the second position in the sentence) is .36 (because the conditional probability of that agent is overall .4, and the probability of the sentence being an active sentence such that the agent occurs in the second position is .9; see Suppl. Fig. 12a). The conditional probability of the action (at the third position) is 1 because the actions are determined by the situations (see Supplementary Note 6 for the rationale behind this predictive relationship between the situation and the action). The conditional probability of the objects (at the fourth position) is either .7 (for high probability objects) or .3 (for low probability objects) so that it is .5 on average, and the conditional probability of the location (at the fifth position) is 1 because the locations are determined by the objects. Thus, the constituents’ conditional probabilities do not gradually decrease across the course of the sentences.

The finding that semantic update nonetheless gradually decreased over successive words in these sentences (see Fig. 2c and Supplementary Fig. 1c) suggests that the SG layer activation does not perfectly track conditional probabilities. Even if an incoming word can be predicted with a probability of 1.0 so that an ideal observer could in principle have no residual uncertainty, the presentation of the item itself still produces some update, indicating that the model retains a degree of uncertainty, consistent with the ‘noisy channel’ model[11]. In this situation, as we should expect, the SG anticipates the presentation of the item more strongly as additional confirmatory evidence is accumulated, so that later perfectly predictable constituents are more strongly anticipated than earlier ones. In summary, the model’s predictions reflect accumulation of predictive influences, rather than completely perfect instantaneous sensitivity to probabilistic constraints in the corpus.

Supplementary Note 6

Aspects of the model environment relevant to the first reversal anomaly simulation (Simulation 9). In the model environment, the situations predict specific actions with a probability of 1. This prevented the critical words (i.e., the actions) from being much better predictable in the reversal anomaly condition where they are preceded by objects (which in the model environment also predict specific actions with a probability of 1) as compared to the congruent condition where they are preceded by agents (which are not predictive of specific actions at all). Of course, situations do not completely determine actions in the real world. However, the rationale behind the decision to construct the corpus in that way to simulate the reversal anomaly experiment by Kuperberg and colleagues (2003)[1] was that the range of plausibly related actions might be similar for specific situations and specific objects such that actions are not much better predictable in the reversal anomaly than in the congruent condition. A relevant difference between both conditions was that in the reversal anomaly condition the model initially assumed the sentences to be in passive voice, because during training, sentences with the objects presented before the actions had always been in passive voice (see Supplementary Fig. 12a). Thus, when the critical word was presented without passive marker (i.e., “by”), the model revised its initial assumptions in that regard in the reversal anomaly condition while there was no need for revision in the congruent condition.

An oversimplification contained in this simulation is that the model never experiences the eggs in any other role than the patient role, even though eggs can occupy other roles in real language environments such as in the sentence “At breakfast, the eggs ruined the omelet”.

This shortcoming is addressed in the simulation of another type of reversal anomaly (Simulation 11).

Supplementary Note 7

Conditional probabilities of semantic features associated with words presented in changed and normal word order (Simulation 12). When changing position of action and patient (type (1)), the conditional probability of the semantic features associated with the critical word (not at this position in the sentence but in general within the described event) is .7 in the condition with the changed word order and 1.0 in the condition with the normal word order. When changing position of agent and action (type (2)), the conditional probability of the semantic features associated with the critical word (again, crucially, not at this position in the sentence but in general within the described event) is 1.0 in the condition with the changed word order and .4 in the condition with the normal word order. Thus, while changes in word order also entail changes in the amount of semantic update of event features, the design of the simulation ensures that influences of word order (syntax) and semantic update can be dissociated. Specifically, the surprise concerning the semantic features of the described event was on average .15 in the condition with the changed word order (.3 for type (1) and 0.0 for type (2)) while it was on average .3 in the condition with the normal word order (0.0 for type (1), and .6 for type (2)).

Supplementary Methods

Supplementary Methods 1

For simulation 10 (e.g. "De speer heft de atleten geworpen", lit: "The javelin has the athletes thrown", relative to "De speer werd door de atleten geworpen", lit: "The javelin was by the athletes thrown")[3], we used the same stimuli as for the first reversal anomaly simulation (Simulation 9), but with Dutch word order. This makes the sentence structures relevant to examining whether the same mechanism that allows the model to account for the small N400 effects in reversal anomaly sentences reported by Kuperberg et al.[1] would also hold when the verb occurs at the end of the sentence. Thus, the relevant experimental conditions contained sentences such as "The pine was by the man *watered*." (i.e., "The pine was watered by the man." with Dutch word order; congruent condition), "The pine has the man *watered*." (i.e. "The pine has watered the man." with Dutch word order; reversal anomaly condition) and "The pine was by the man *drunken*." (i.e., "The pine was drunken by the man." with Dutch word order; incongruent condition).

We trained the model on the same training environment used in the main simulations (see Methods and Supplementary Fig. 12), with the following modifications. The sentence structures were adjusted such that active sentences were changed from e.g., "The man waters the pine." to "The man has the pine watered." and passive sentences were changed from "The pine was watered by the man." to "The pine was by the man watered." We added an additional input unit representing "has" and used a single unit for "was by" because both words now always occurred in direct succession (e.g., "… was by the man watered." instead of "… was watered by the man."). Apart from these adjustments, all parameters of the model and training were kept the same. This implementation does not completely correspond to the empirical experiment[3] in that in our simulation there was no specific relationship between the agent and the action (i.e., the man in the model environment is equally likely to perform all 12 actions and thus was equally likely to water something as he was to drink something, for instance) while in the stimulus material of the empirical experiment there was a specific probabilistic relationship between the agents and the actions (i.e., athletes might be more likely to throw something than to summarize something). However, important for current purposes, this implementation allowed to test whether the way the model accounts for the small N400 increase in reversal anomalies would be robust to changes in word order, i.e. the presentation of two noun phrases prior to the presentation of the verb.

There were eight items in each experimental condition, and SU was computed as the difference in SG layer activation between the third constituent ("man", word *n*-1) and the fourth constituent (the action, word *n*). The results are displayed in Supplementary Figure 4. Consistent with the experimental findings, the SU at the verb in the reversal anomaly sentences is only slightly larger than the SU in the congruent control condition, and the SU for the incongruent verb condition is much larger.

Supplementary Methods 2

Here, we describe the changes to the standard training environment (see Methods and Supplementary Fig. 12) that we implemented to simulate reversal anomalies with two animate event participants that can both occur as agents[2] (Simulation 11). We increased the percentage of passive sentences in the model's environment from 10% to 30% (the implementation of the retrieval-integration model used 50% passive sentences[12]). We do not assume that there are more passive sentences in Dutch than in English, but take the increase of the rate of passive sentences to be a simple approximation to the situation that a major grammatical difference between English and Dutch lies in the number of permissible word orders so that word order is a less reliable cue to meaning in Dutch. Specifically, a study found SV word order to be a valid cue to the agent role in 95/100 of sentences in English but only 35/100 sentences in Dutch[13]. As we assume that the model simultaneously uses all available constraints to map from incoming words to sentence meaning, we assume this variability in terms of word order in Dutch also plays a role in its interpretation of reversal anomaly sentences with two noun phrases that can both be agents. Although the empirical experiment[2] used embedded clauses (e.g., lit: "The fox that on the poacher hunted…"; paraphrase: "The fox that hunted the poacher…"), we used single clause sentences with Dutch word order (e.g., lit: "The fox on the poacher hunted."; paraphrase: "The fox hunted the poacher.").

To capture the relevant features, we extended the training environment for this simulation by adding eight analogous scenarios to the main simulation environment (Supplementary Fig. 5). To keep the overall size of the training environment roughly similar to the main simulation, we eliminated six actions with their respective objects and situations from the environment (Supplementary Fig. 12) when adding these new scenarios. In keeping with the characteristics of the materials used in the experiment, which were often built around a typical event that quickly comes to mind when particular participants are involved (for example, surgeons typically operate on patients), we constructed each of the new scenarios around a probable event involving a central agent doing a central action to a central patient (e.g., the poacher hunting the fox or the surgeon operating on the patient), but alternative events can occur with lower probability as well. In particular, the central patient can also perform the central action, but not towards the central agent. For instance, the fox can also hunt (though not the poacher) and the patient can also cut into something (though not into the surgeon). Furthermore, there are alternative less specific actions as well (such as approaching, watching, standing or sitting in front of) that can be performed by all sorts of agents (including the central patients) towards all sorts of patients (including the central agents). Thus, the central patients can sometimes also be agents in events involving the central agents, e.g., the fox can watch the poacher and the patient can stand in front of the surgeon.

There is considerable variability among the materials used in the empirical experiment[2] as is apparent from the different examples involving the fox/poacher and patient/ surgeon, and the ERP is averaged across all these materials. Thus, we designed our scenarios (see Suppl. Fig. 5 for details) based on our examination of the entire set of the experimental materials, which we obtained from the authors, instead of trying to capture any particular scenario exactly. Furthermore, we cannot claim to have exactly matched the average probabilities of actions performed by the various participants across the full set of materials used in the actual experiment. The richness and diversity of the experimental materials along a number of relevant dimensions makes it difficult to determine how well we have approximated the factors that influence the construction of a representation of meaning in these sentences. The current scenarios thus provide a proof of concept that the model can capture the empirical data when taking into account the elements of reversal anomalies described above. This proof-of-concept approach in light of the complexity of the issue is somewhat similar to the approach taken in implementing the training environment for the retrieval integration model[12].

Apart from these changes, all parameters of the model and training were kept the same, and again, 10 instances of the model were trained on 800,000 sentences each.

Supplementary Methods 3

We trained a classic simple recurrent network[14] consisting of an input and output layer with 74 units each, as well as a hidden and context layer with 100 units each, on the same standard training corpus as the SG model (see Supplementary Fig. 12a). Except for the architectural difference, all parameters were kept the same. We then simulated influences of violations of word order, reversal anomalies, and development, as described for the SG model (see *Methods/ Simulation of empirical findings*). The measure for surprisal that we set in relation to N400 amplitudes consists in the summed magnitude of the cross-entropy error induced by the current word (word *n*).

Supplementary Methods 4

All reported statistical results are based on ten runs of the model each initialized independently (with initial weights randomly varying between +/- .05) and trained with independently-generated training examples as described in section Methods/ Environment (*n*=800,000, unless otherwise indicated). In analogy to subject and item analyses in empirical experiments, we performed two types of analyses on each comparison, a model analysis with values averaged over items within each condition and the 10 models treated as random factor, and an item analysis with values averaged over models and the items (*n* ranging between 8 and 16; please see section Methods/ Simulation of empirical findings, and figure captions for the exact number of items in each simulation experiment) treated as random factor. There is much less noise in the simulations as compared to empirical experiments such that the relatively small sample size (10 runs of the model and 8 to 16 items per condition) should be sufficient.

We used two-sided paired t-tests to analyze differences between conditions; when a simulation experiment involved more than one comparison, significance levels were Bonferroni-corrected within the simulation experiment. Effect size (Cohen's d) was computed using the function computeCohen_d in MATLAB. To test for the interaction between repetition and congruity, we used a repeated measures analysis of variance (rmANOVA) with factors Repetition and Congruity. To analyze whether our data met the normality assumption for these parametric tests, we tested differences between conditions (for the t-tests) and residuals (for the rmANOVA) for normality with the Shapiro-Wilk test. Using study-wide Bonferroni correction to adjust significance levels for the multiple performed tests, results did not show significant deviations from normality (all *p*s > .15 for the model analyses and > .32 for the item analyses) except for the item analysis of the change in word order which showed a marginally significant effect (Simulation 12; Supplementary Fig. 2c; *p* = .066; this might be due to the items in this simulation experiment consisting of two types with slightly different characteristics; see section *Methods/ Simulation of empirical findings* and Supplementary Note 7). This effect did not change when using the Wilcoxon signed rank test, which does not depend on the normality assumption. Specifically, the effect of the change in word order (Simulation 12) in the item analysis did not reach significance neither in the t-test (see caption of Supplementary Fig. 2c) nor in the Wilcoxon signed rank test (*p* = .10). To further corroborate our results we additionally tested all comparisons with deviations from normality at uncorrected significance levels <.05 using the Wilcoxon signed rank test; all results remained significant. Specifically, in the model analyses deviations from normality at uncorrected significance levels were detected for the semantic incongruity effect (Simulation 1; Fig. 2a; *p* = .043), the frequency effect (Simulation 5; Fig. 2e; *p* = .044), the difference between categorically related incongruities and congruent completions (Simulation 4; Fig. 2d; *p* = .0053), as well as the difference between congruent and reversal anomaly sentences (Simulation 11; Fig. 3b; *p* = .010) and between congruent and incongruent sentences (Simulation 11; Fig. 3b; *p* = .013) in the third reversal anomaly simulation. Wilcoxon signed rank tests confirmed significant effects of semantic incongruity (Fig. 2a; *p* = .002), lexical frequency (Fig. 2e; *p* = .037), a significant difference between categorically related incongruities and congruent sentence continuations (Fig. 2d; *p* = .002), as well as significant differences between congruent and reversal anomaly sentences (Fig. 3b; *p* = .002) and between congruent and incongruent sentences (Fig. 3b; *p* = .002) in the third reversal anomaly simulation. In the item analyses, deviations from normality at an uncorrected significance level were detected for the difference between low constraint unexpected endings and expected endings (Simulation 13; Supplementary Fig. 2d; *p* = .03), the difference between incongruent completions and reversal anomalies in the first reversal anomaly simulation in the SG model (Simulation 9; Supplementary Fig. 2a; *p* = .012) as well as in the SRN (Supplementary Fig. 10a; *p* = .043), and for the difference between changed and normal word order in the SRN (Supplementary Fig. 10b; *p* = .011). Again, Wilcoxon signed rank tests confirmed significant differences between low constraint unexpected endings and expected endings (Supplementary Fig. 2d; $p = .002$), between the incongruent completions and reversal anomalies in the SG model (Supplementary Fig. 2a; $p = .0078$) and the SRN (Supplementary Fig. 10a; $p = .039$), as well as a significant influence of word order in the SRN (Supplementary Fig. 10b; $p < .001$).

Using Levene's test, we detected violations of the assumption of homogeneity of variances (required for the rmANOVA used to analyze the interaction between repetition and congruity; Fig. 6 and Supplementary Fig. 9) in the item analysis, $F_2(3) = 12.05, p < .001$, but not in the model analysis, $F_1(3) < 1$. We nonetheless report the ANOVA results for both analyses because ANOVAs are typically robust to violations of this assumption as long as the groups to be compared are of the same size. However, we additionally corroborated the interaction result from the item ANOVA by performing a two-tailed paired t-test on the repetition effects in the incongruent versus congruent conditions, i.e. we directly tested the hypothesis that the size of the difference in the model's N400 correlate between the first presentation and the repetition was larger for incongruent than for congruent sentence completions: incongruent (first – repetition) > congruent (first – repetition). Indeed, the size of the repetition effects significantly differed between congruent and incongruent conditions, $t_2(9) = 10.99, p < .001$, and the differences between conditions did not significantly deviate from normality, $p = .44$, thus fulfilling the prerequisites for performing the t-test.

In general, systematic deviations from normality are unlikely for the results by-model (where apparent idiosyncrasies are most probably due to sampling noise), but possible in the by-item data. Thus, while we present data averaged over items in the figures in the main text in accordance with the common practice in ERP research to analyze data averaged over items, for transparency we additionally display the data averaged over models as used for the by-item analyses (see Supplementary Fig. 1, 2, 4, 7-10).

Supplementary Discussion

Implicit probabilistic theory of meaning

The theory of meaning embodied in the Sentence Gestalt model holds that sentences constrain an implicit probabilistic representation of the meanings speakers intend to convey through these sentences. The representation is implicit in that no specific form for the representation is prescribed, nor are - in the general form of the theory - specific bounds set on the content of the representation of meaning. (In any specific implementation of the theory, the content of the representation of meaning is prescribed by the range of possible probes and queries, which in the case of our implementation correspond to the vectors encoding the pairs of thematic roles and their fillers of described events; see below). Sentences are viewed as conveying information about situations or events, and a representation of meaning is treated as a representation that provides the comprehender with a basis for estimating the probabilities of aspects of the situation or event the sentence describes. To capture this we characterize the ensemble of aspects as an ensemble of queries about the event, with each query associated with an ensemble of possible responses. The query-answer form is used instead of directly providing the complete event description at the output layer to keep the set of probes and fillers more open-ended and to suggest the broader framework that the task of sentence comprehension consists in building internal representations that can be used as a basis to respond to probes[15]. In the general form of the theory, the queries could range widely in nature and scope, encompassing, for example, whatever the comprehender should expect to observe via any sense modality or subsequent linguistic input, given the input received so far. This includes queries supporting expectations concerning the content of stative sentences (e.g., "Her hair is red.") and probes supporting the anticipation of aspects of meaning of questions or commands based on representations concerning the current state of knowledge and intentions of the speaker, etc. For instance, the question "Where is the bathroom?" communicates that the speaker would like to know the location of the bathroom, and the command "Please close the door." communicates that the speaker wants the listener to close the door. Thus, it is important to note that even though the current implementation focuses on sentences describing events, the theory is thought of as applying to language comprehension in general. In implementations to date, at least four different query formats have been considered[7,16,17], including a natural language-based question and answer format (Fincham & McClelland, 1997, Abstract). Queries may also vary in their probability of being posed (hereafter called *demand probability*), and the correct answer to a particular query may be uncertain, since sentences may be ambiguous, vague or incomplete. An important aspect of the theory that receives little attention in many other theories of sentence comprehension is that aspects of meaning can often be estimated without being explicitly described in a sentence, due to knowledge acquired through past experience[7]. If events involving cutting steak usually involve a knife, the knife would be understood, even without ever having been explicitly mentioned in a sentence.

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

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

More formally, the learning environment is treated as producing sentence-event-description pairs according to a probabilistic generative model. The sentence consists of a sequence of words, while the event-description consists of a set of queries and associated responses. Each such pair is called an *example.* The words in the sentence are presented to the neural network in sequence, and after each word, the system can be probed for its response to each query, which is conditional on the words presented so far (we use w_n to denote the sequence of words up to and including word n). The goal of learning is to minimize the expected value over the distribution of examples of a probabilistic measure (the Kullback-Leibler divergence, D_{KL}) of the difference between the distribution of probabilities p over possible responses r to each possible query and the model's estimates ρ of the distribution of these probabilities, summed over all of the queries q occurring after each word, and over all of the words in the sentence. In this sum, the contribution of each query is weighed by its demand probability conditional on the words seen so far, represented $p(q|w_n)$. We call this the *expected value E of the summed divergence measure*, written as:

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

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

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

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

Because the real learning environment is rich and probabilistic, the number of possible sentences that may occur in the environment is indefinite, and it would not in general be possible to represent the estimates of the conditional probabilities explicitly (e.g. by listing them in a table). A neural network solves this problem by providing a mechanism that can process any sequence of words and associated queries that are within the scope of its environment, allowing it to generate appropriate estimates in response to queries about sentences it has never seen before[7].

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

The implemented query-answer format and standard network learning rule

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

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

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

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

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

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

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

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

Now, the network uses units whose activation *a* is given by the logistic function of its net input, such that $a = 1/(1 + e^{-net})$, where the net input is the sum of the weighted influences of other units projecting to the unit in question, plus its bias term. As has long been known[18], the negative of the gradient of this cross-entropy measure with respect to the net input to the unit is simply $t(f|C_e) - a(f|C_e)$. This is the signal back-propagated through the network for each feature in each context during standard network training (see section *Methods/ Training protocol* for more detail).

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

Others have proposed probabilistic measures of the surprise produced by perceptual or linguistic inputs[19,20]. In the framework of our approach to the characterization of sentence meaning, we adapt one of these proposals[19], and use it to propose measures of three slightly different conceptions of surprise: The normative surprise, the subjective explicit surprise, and the implicit surprise – the last of which corresponds closely to the measure we use to model the N400.

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

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

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

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

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

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

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

Our third measure, the implicit surprise (IS) is a probabilistically interpretable measure of the change in the pattern of activation over the learned internal meaning representation (corresponding to the SG layer in the model). Since the unit activations are constrained to lie in the interval between 0 and 1, they can be viewed intuitively as representing estimates of probabilities of implicit underlying meaning dimensions or *microfeatures*[21] that together constrain the model's estimates of the explicit feature probabilities. In this case we can define the implicit surprise as the summed KL divergence between these implicit feature probabilities before and after the occurrence of word *n*, using a_i to represent the estimate of the probability that the feature characterizes the meaning of the sentence and $(1 - a_i)$ to represent the negation of this probability:

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

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

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

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

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

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

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

Semantic update driven learning rule

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

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

A single {*sentence, event*} pair chosen from the environment would then provide a sample from this distribution. As is the case in the standard training regime, the negative of the gradient with respect to the net input to a given output feature unit in the query network after a given probe is simply $t(f|q,e) - a(f|q,w_N)$. This is then the error signal propagated back through the network. To train the network to make better estimates of the feature probabilities from the next to last word in the sentence (word *N*-1), we can use the difference between the activations of the output units after word *N* as the teaching signal for word *N*-1, so for a given feature unit the estimate of the gradient with respect to its net input simply becomes $a(f|q,w_N) - a(f|q,w_{N-1})$. Using this approach, as $a(f|q,w_N)$ comes to approximate *t*(*f*|*q*,*e*) it thereby comes to approximate the correct target for *a*(*f*|*q*, *N*-1). This cycle repeats for earlier words, so that as *a*(*f*|*q*, *N*-1) comes to approximate *a*(*f*|*q*, *N*) and therefore *t*(*f*|*q*, *e*) it also comes to approximate the correct teacher for *a*(*f*|*q*, *N*-2), etc. This approach is similar to the temporal difference (TD) learning method used in reinforcement learning[22] in situations where reward becomes available only at the end of an episode, except that here we would be learning the estimates of the probabilities for all of the queries rather than a single estimate of the final reward at the end of an episode. This method is known to be slow and can be unstable, but it could be used in combination with learning based on episodes in which teaching information is available throughout the processing of the sentence, as in the standard learning rule for the SG model.

The semantic update based learning rule we propose extends the idea described above, based on the observation that the pattern of activation over the SG layer of the update network serves as the input pattern that allows the query network to produce estimates of probabilities of alternative possible responses to queries after it has seen some or all of the words in a sentence. Consider for the moment an ideally trained network in which the presentation of each word produces the optimal update to the SG representation based on the environment it had been trained on so far, so that the activations at the output of the query network would correspond exactly to the correct probability estimates. Then using the SG representation after word n+1 as the target for training the SG representation after word n would allow the network to update its implicit representation based on word n to capture changes in the environmental probabilities as these might be conveyed in a sentence. More formally, we propose that changing the weights in the update network to minimize the Implicit Surprise allows the network to make an approximate update to its implicit probabilistic model of sentence meaning, providing a way for the network to learn from linguistic input alone. The negative of the gradient of the Implicit Surprise with respect to the net input to SG unit i after word n is given by $a_i(w_n) - a_i(w_{n-1})$. This is therefore the signal that we back propagate through the update network to train the connections during implicit temporal difference learning. As noted in the main text, the sum over the SG units of the absolute value of this quantity also corresponds to the SU, our model's N400 correlate. The model would not be able to learn language based on this semantic update driven learning rule alone. We assume that language learning proceeds by a mixture of experience with language processed in the context of observed events (as in the standard training regime) and processed in isolation (as with the semantic update driven learning rule), possibly with changing proportions across development. Future modeling work should explore this issue in more detail.

Supplementary References

1. 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).
2. 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).
3. 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).
4. Mikolov, T., Deoras, A., Povey, D., Burget, L. & Cernocky, J. H. Strategies for Training Large Scale Neural Network Language Models. in *IEEE Workshop on Automatic Speech Recognition and Understanding* (2011).
5. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. *Emnlp2014.Org* at <http://emnlp2014.org/papers/pdf/EMNLP2014162.pdf>
6. Delong, K. A., Quante, L. & Kutas, M. Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia* **61,** 150–162 (2014).
7. St. John, M. F. & McClelland, J. L. Learning and applying contextual constraints in sentence comprehension. *Artif. Intell.* **46,** 217–257 (1990).
8. Sanford, A. J. & Sturt, P. Depth of processing in language comprehension: Not noticing the evidence. *Trends Cogn. Sci.* **6,** 382–386 (2002).

9. Ferreira, F., Bailey, K. G. D. & Ferraro, V. Good-Enough Representations in Language Comprehension. *Curr. Dir. Psychol. Sci.* **11,** 11–15 (2002).
10. Staab, J., Urbach, T. & Kutas, M. Negation processing in context is not (always) delayed. *Cent. Res. Lang. Tech. Rep.* **20,** 3–34 (2009).
11. Levy, R. A noisy-channel model of rational human sentence comprehension under uncertain input. in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* 234–243 (2008). doi:10.3115/1613715.1613749
12. 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).
13. McDonald, J. L. The development of sentence comprehension strategies in English and Dutch. *J. Exp. Child Psychol.* **41,** 317–335 (1986).
14. Elman, J. L. Finding Structure in Time. *Cogn. Sci.* **14,** 179–211 (1990).
15. McClelland, J. L., St. John, M. & Taraban, R. Sentence comprehension: A parallel distributed processing approach. *Lang. Cogn. Process.* **4,** 287–336 (1989).
16. Rohde, D. L. T. A Connectionist Model of Sentence Comprehension and Production. (Carnegie Mellon University, 2002).
17. Bryant, B. D. & Miikkulainen, R. *From Word Stream to Gestalt: A Direct Semantic Parse for Complex Sentences*. (2001).
18. Hinton, G. I. Connectionist Learning Procedures. *Mach. Learn. -- an Artif. Intell. Approach* **III,** 555–610 (1990).
19. Itti, L. & Baldi, P. Bayesian Surprise Attracts Human Attention. 1–8 (2006). doi:10.1016/j.visres.2008.09.007
20. Levy, R. Expectation-based syntactic comprehension. *Cognition* **106,** 1126–1177 (2008).
21. Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. in *Parallel Distributed Processing* (eds. Rumelhart, D. E. & McClelland, J. L.) 77–109 (MIT Press, 1986). doi:10.1146/annurev-psych-120710-100344
22. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*. (MIT Press, 1998).