PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Opinion piece

Cite this article: Rabovsky M, McClelland JL. 2019 Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension. *Phil. Trans. R. Soc. B* **375**: 20190313.
http://dx.doi.org/10.1098/rstb.2019.0313

Accepted: 16 October 2019

One contribution of 16 to a theme issue 'Towards mechanistic models of meaning composition'.

Subject Areas:
cognition, neuroscience

Keywords:
language, meaning, event-related brain potentials, neural networks, N400, P600

Authors for correspondence:
Milena Rabovsky
e-mail: rabovsky@uni-potsdam.de
James L. McClelland
e-mail: jlmcc@stanford.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4723004.

THE ROYAL SOCIETY
PUBLISHING

Quasi-compositional mapping from form to meaning: a neural network-based approach to capturing neural responses during human language comprehension

Milena Rabovsky[1] and James L. McClelland[2]

[1]Department of Psychology, University of Potsdam, Karl-Liebknecht-Strasse 24-25, 14476 Potsdam, Germany
[2]Department of Psychology, Stanford University, 450 Jane Stanford Way, Stanford, CA 94305, USA

MR, 0000-0001-7729-1027; JLM, 0000-0002-8217-405X

We argue that natural language can be usefully described as quasi-compositional and we suggest that deep learning-based neural language models bear long-term promise to capture how language conveys meaning. We also note that a successful account of human language processing should explain both the outcome of the comprehension process and the continuous internal processes underlying this performance. These points motivate our discussion of a neural network model of sentence comprehension, the Sentence Gestalt model, which we have used to account for the N400 component of the event-related brain potential (ERP), which tracks meaning processing as it happens in real time. The model, which shares features with recent deep learning-based language models, simulates N400 amplitude as the automatic update of a probabilistic representation of the situation or event described by the sentence, corresponding to a temporal difference learning signal at the level of meaning. We suggest that this process happens relatively automatically, and that sometimes a more-controlled attention-dependent process is necessary for successful comprehension, which may be reflected in the subsequent P600 ERP component. We relate this account to current deep learning models as well as classic linguistic theory, and use it to illustrate a domain general perspective on some specific linguistic operations postulated based on compositional analyses of natural language.

This article is part of the theme issue 'Towards mechanistic models of meaning composition'.

1. Introduction

Language ultimately aims to convey meaning, but despite the crucial role of meaning, the mechanistic basis of the processing of meaning in language remains incompletely understood. One much debated issue is the composition of meaning, i.e. the construction of an integrated meaning representation. In sentence comprehension, this corresponds to the formation of a representation of sentence meaning based on the sequence of individual words.

Correspondingly, in the world, situations and events are composed of entities and actions, each contributing to the overall meaning. The situations and events are often composed according to a specific structure or regularity in that they consist of specific roles entailing specific expectations, which can be filled by different event participants compatible with these expectations. For instance, in a restaurant, there are roles such as waiters and customers, or in a family, there are roles such as father, mother and children, each of which can be filled by a large variety of people. Even more interchangeable event participants can take on roles in events involving actions such as giving, taking, buying, etc.

A widely discussed concept in investigations of the formation of an integrated representation of sentence meaning is the concept of compositionality. We are deeply interested in the neurocognitive processes underlying the formation of an integrated meaning representation during language comprehension, but are not experts in the linguistic and philosophical debates concerning the concept of compositionality. To set our ideas in relation to these debates, we consulted the literature on compositionality, with basic chapters on compositionality written by experts in the field as our entry points.

A classic definition of compositionality is given by Partee *et al.* [1], who define the principle of compositionality as the statement that 'the meaning of a complex expression is a function of the meanings of its parts and of the syntactic rules by which they are combined' [1, p. 318]. Pelletier notes in the Oxford Encyclopedia chapter on semantic compositionality [2, p. 1] that 'most linguists have heard of semantic compositionality. Some will have heard that it is the fundamental truth of semantics. Others will have been told that it is so thoroughly and completely wrong that it is astonishing that it is still being taught'. An interesting perspective on compositionality is also provided by Janssen in his chapter on compositionality in the Handbook on Logic & Language [3] where he argues that 'the principle [of compositionality] should not be considered an empirically verifiable restriction, but a methodological principle' [3, p. 419]. Indeed, a common topic of discussion and research concerning compositionality seems to consist in the raising of counterexamples (i.e. sentences that seem incompatible with the principle) and the subsequent search for a solution, which allows for a compositional analysis of the sentence by extending the concept of parts, syntactic rules or both, as given in the definition of compositionality above (see [3] for many such counterexamples and their compositional solutions).

Based on this exercise, a viable position seems to be that any natural language fragment can be analysed compositionally as long as enough complexity is built into the lexicon, syntax or both. As an example, consider the meaning of 'she felt the baby kick' and 'he felt the rifle kick', which describe very different events. We would not capture the understood meaning of either sentence by simply placing a representation of the meaning of each of the words into corresponding places in compositional structures, as the nature of the action expressed by the word 'kick' is very different in the two sentences above. For example, in the first sentence, we understand the baby's kick as an occurrence that may be a mother's first experience of the movement of her unborn baby. This seems to suggest that the understood meaning of the sentence is incompatible with compositionality. However, it is argued that complexity added to the lexicon, the syntax or both, will allow for a compositional analysis of such sentences despite the initial apparent incompatibility. For example, the entries for 'rifle' and 'baby' could select specific definitions of the meaning of 'kick'. In the light of this flexibility, we agree with Janssen who suggests that 'the real question is not whether a certain phenomenon can be analysed compositionally, but what makes the overall theory (un)attractive or (un)acceptable' [3, p. 441]. Relatedly, he notes that it has been suggested that the necessity to stretch the concepts of the lexicon and syntax to achieve compositionality makes compositionality 'a vacuous principle' [3, p. 457]. Jannsen goes on to reject this view, because from his perspective, 'the challenge of compositional semantics is not to prove the existence of such semantics, but to obtain one'. Here, as noted above, compositionality becomes a methodological principle allowing researchers to 'design a function that assigns meanings'. Other researchers do view compositionality not as a method but as a factual claim, open to empirical testing [4]. For those who see compositionality as a factual claim, a possible empirical research programme (related to the question raised in the next paragraph) is to investigate whether the operations postulated to enable compositional analyses of natural language sentences are reflected in online measures of human language comprehension [5–9].

Indeed, an important distinction we have not yet considered is the distinction between natural language understanding, defined as the human process of understanding natural language, and natural language semantics, defined as the model-theoretical analysis of natural language sentences in terms of their truth values. It is possible that compositionality holds to a different extent in these two cases. Partee, who coined the classic definition of compositionality, argues (according to Janssen [3]), that a finite complete compositional semantics that really deals with human natural language understanding is not possible [10,11]. Specifically, Janssen summarizes Partee's view as follows [3, p. 447]: 'Compositional model-theoretic semantics is possible and important, but one should understand the limits of what it can do. In a system of compositional semantics the flexibility of language is abstracted away. Therefore it is too rigid to describe the real life process of communication, and limits the description of language users to creatures or machines whose minds are much more narrowly and rigidly circumscribed than those of human beings. This underscores the argument that a theory of natural language semantics should be distinguished from a theory of natural language understanding'. We agree with this view.

We respect that other researchers may have different views on this issue and we agree that it can be a valuable research programme to investigate in how far concepts from model-theoretic compositional semantics can account for aspects of human language comprehension as noted above (see §5 'A neural network perspective on linguistic operations postulated in compositional analyses' for further discussion). At the same time, we agree with Partee that restricting theories of human language understanding to lie within what is possible when adhering to her definition of compositionality places undesirable limits on the development of a theory of human natural language understanding, which is our main interest.

Elsewhere, we have described natural language as well as natural objects and events as quasi-compositional or quasi-regular [12–14]. We use the phrase 'quasi-compositional' to contrast our view with the usage of compositional found in Fodor & Pylyshyn [15]. They give 'compositionality' a restricted meaning, in which the contribution each lexical item makes to the meaning of a sentence is always the same. They appealed to such a restricted definition of compositionality, arguing that it allows language to be productive. On p. 42, they wrote 'in fact, you need a further assumption, which we'll call the 'principle of compositionality': insofar as a language is systematic, a lexical item must make approximately the same semantic contribution to each expression in which it occurs. It is, for example, only insofar as 'the' 'girl', 'loves' and 'John' make the same semantic contribution to 'John loves the girl' that they make to 'the girl loves John' that understanding the one sentence implies understanding the other'. However, as already discussed, such a restricted form of compositionality does not fully capture the mapping from language to meaning, and in our view this explains why compositional approaches have (to the best of our knowledge) not yet proven successful in large-scale machine approaches to natural language processing (NLP).

Rather than adding complexity to try to address the wide range of cases where the restricted form of compositionality fails, we instead note that there is generally a degree of compositionality coupled with a degree of specificity in every linguistic or conceptual object, from phonetic objects in spoken words to conceptual objects in events and situations [12–14]. In our 'felt/kick' examples, there is a force exerted by an object on a target sentient being by either a rifle or a baby causing a sensation in the target. We seek a system that captures this core (the regular or compositional part) and allows it to be available for generalization, just as Fodor and Pylyshyn desire, while simultaneously providing a means to capture the idiosyncratic elements that go beyond it (thereby making the compositionality only partial). We have the hope that our approach can capture the flexibility of language Partee mentioned in the passage quoted above,[1] while also capturing the regular or systematic part of language as just described.

Specifically, we adopt neural network models, which are not based on pre-defined symbols and explicit rules accompanied by lists of exceptions, but rather learn continuous nonlinear functions mapping from inputs to outputs without imposing constraints on the internal representation used. We adopt neural networks because we believe they hold promise to capture both the regular or systematic as well as the exceptional and idiosyncratic aspects of language, of our construals of the world, and of the mapping between them, and that the human process of learning to understand language consists in learning functions mapping between these domains. These learned functions are more flexible than strict rules, extracting general statistical regularities, which can yield rule-like behaviour, while being at the same time sensitive to specific information, allowing them to capture exceptions as well as graded influences of both general and specific information as in quasi-regularities (e.g. 'kick' in the examples above). Here simplicity is achieved by letting the gradient-based learning process that adjusts the connection weights in the neural network sort out both the systematic and idiosyncratic aspects of the mapping, rather than attempting to capture them in explicit form.

Technical developments, increases in computing power, and the availability of large corpora of training data have resulted in deep neural network models making great progress in large-scale NLP such as in machine translation [16] as well as natural language understanding tasks as assembled in the General Language Understanding Evaluation benchmark [17]. Considerable challenges remain for neural network models in the domain of language comprehension (see e.g. [18,19]) and there is certainly room for improvement [20]. However, we are encouraged by the current rate of progress on problems in NLP such as capturing scope phenomena with negations [21] and quantifiers [22], the ability to use a new word in combination with other words after learning about it when it only occurs by itself [23], as well as a broad range of other NLP tasks [24–27]. Together with the struggles of rule-based approaches in NLP, these developments seem to indicate there is certainly reason to continue exploring neural network models as alternatives to more explicit, rule-based compositional approaches when aiming to obtain functions mapping from natural language to understood meaning.

Against this backdrop, our work asks whether and how neural network models can account for neural responses observed during human language comprehension. In our view, a convincing theory of human language comprehension needs to address not only successful comprehension as measured by behavioural responses, but also the continuous internal processes underlying this performance, operating closer to what Marr called the 'algorithmic' rather than 'computational' level [28]. A pertinent measure of online language comprehension processes is provided by event-related brain potentials (ERPs), which offer time-resolved measures of electrical brain activity during comprehension. While behavioural responses in specific tasks are typically influenced by a mixture of processes, specific ERP components (i.e. peaks and valleys in the waveform which are reproducible in terms of their latency, polarity, topographical distribution and functional characteristics) reflect specific sub-processes taking place between stimulus presentation and response, making it possible to disentangle and investigate these sub-processes separately.

Here, we mainly focus on the N400 component, which has been used in more than a thousand empirical studies and seems to reflect the initial, relatively automatic brain response to incoming language input [29]. We will argue that N400s are currently most comprehensively accounted for by a neural network model of sentence comprehension called the Sentence Gestalt (SG) model [30], which learns a function mapping from the sequence of incoming words to a representation of the event or situation the sentence describes. In this model, meanings of components (words and phrases) are not assembled into an overall meaning representation. Instead, they provide clues to meaning [31], in that each incoming word constrains the representation of all of the objects, actions and relationships characterizing the described event.

We see the SG model as capturing the N400, which we view (as noted above) as reflecting as relatively automatic brain response. This process may not always result in a coherent interpretation and in these cases, cognitive control may be necessary for successful comprehension. We argue that control processes and/or their outcome are reflected in another ERP component, the P600, which sometimes follows the N400 and seems more dependent on attention and task variables [32]. The distinction between automatic and controlled processes or related distinctions such conscious versus unconscious perception, and goal-directed versus habitual control in decision making are ubiquitous in cognitive and systems neuroscience. We suggest that this distinction plays an important role in language comprehension as well. Sentence comprehension often seems like an effortless process—it seems almost difficult not to understand simple sentences (e.g. 'the boy kicked the ball') when hearing or reading them. However, when asked for instance whether the sentence 'The dog was bitten by the man' describes a plausible event, there might be an initial moment of confusion and some effort needed to reach the conclusion that the described event is highly implausible. We suggest that consideration of both automatic and controlled aspects of language comprehension can help us understand the internal processes involved in human sentence comprehension and thus the functional basis of language-related ERPs such as the N400 and P600. This dimension currently seems to be missing in modern neural network models of language comprehension in use by the machine learning community, where the goal is to solve applied problems. However, the electrophysiology of language comprehension offers clues that we argue must be attended to if we are to fully understand how the human brain processes language.

Below, we first describe our account of language-related ERPs, based on our model, which was initially developed [33] to capture the quasi-compositional aspects of language comprehension. We discuss aspects of the model's relation to modern deep learning models and to linguistic theory. We also consider the P600, which, as we have suggested above, we take to reflect more-controlled processes that occur when the more automatic process reflected in the N400 results in conflict or uncertainty. Finally, we use our model to illustrate a domain general perspective on some specific linguistic operations postulated based on compositional analyses of natural language.

2. Modeling the N400 component of the event-related brain potential

As just noted the N400 is of particular relevance to meaning construction in the brain. The N400 is a relative negativity at centro-parietal electrode sites peaking around 400 ms after the presentation of a potentially meaningful stimulus (with 'meaning' broadly construed to include the meaning of mathematical formulas, specific sounds, etc.). The N400 is reliably modulated by a wide variety of semantic variables, e.g. at the level of single words, word pairs, sentences and discourse as well as by semantic factors in non-linguistic domains such as pictures, sounds, maths, etc. The first observation was that N400s are reduced for words that are semantically expected as compared to words that are implausible or unexpected in a given context. However, N400s are also reduced for words following semantically or associatively related prime words, for repeated words, and for words of higher lexical frequency (see [20] for review). Interestingly, N400 effects have been observed even during the attentional blink, a paradigm resulting in an inability to report the presented stimulus [34], suggesting that the process underlying N400s occurs relatively automatically as discussed above.

In recent years, there has been growing interest in linking N400s to neural network models [30,35–41]. Our own account focuses on modelling N400 amplitude, i.e. its magnitude, at a functional level, thus abstracting away from the spatio-temporal profile of the physiological response. Our account is based on the observation that N400s seem to crucially depend on the discrepancy between probabilistically expected and encountered meaning, which can be seen as corresponding to an implicit prediction error. From a neural network modelling perspective [42], processing and learning is based on the generation of implicit expectations based on experienced statistical regularities in the environment, reflected in model-generated activation at the output layer. This model-generated activation is compared to actual observations (implemented as the correct target activation). The implicit prediction error, i.e. the difference between implicit expectations and actual observations (i.e. between model generated and correct activation) serves as the learning signal driving adaptation and learning, implemented as the adaptation of connection strengths in the network via error-backpropagation [43]. Based on these assumptions, Rabovsky & McRae [38] simulated N400s as the network error in a neural network model of word meaning in a series of simulations, which they took to suggest that N400 amplitudes reflect an implicit prediction error and learning signal in the semantic system.

However, this model left open how the presumed prediction error is implemented in neural activation. This issue was addressed by Rabovsky *et al*. [30], who extended the approach by Rabovsky & McRae [38] to sentence processing and refined the notion of the assumed implicit prediction error in crucial ways. Specifically, N400s were simulated as the stimulus-induced change in activation at a hidden 'Sentence Gestalt' layer [33], which implicitly and probabilistically represents predicted sentence meaning. Because the activation pattern at any given point in sentence presentation corresponds to the model's implicit prediction of all the semantic features involved in the described event, the change in this activation state induced by the new incoming stimulus corresponds to the implicit prediction error contained in the previous representation. Here, implicit semantic prediction error is dynamically implemented as the change in an internal probabilistic representation of meaning. Using this approach, Rabovsky *et al*. successfully simulate 16 distinct and diverse empirically observed N400 effects, including N400 effects at the sentence level (semantic violations, categorically related semantic violations, cloze probability, a word's position in the sentence) and word level (semantic, associative and repetition priming, lexical frequency). In addition, they simulate several N400 effects related to language learning and development (N400 effects across development, in a newly learned language, interaction between repetition and semantic violations). Moreover, they demonstrate the specificity of the model's N400 correlate by showing that it is not sensitive to several factors that do not influence the N400. Factors that do *not* influence the N400 include reversal anomalies (see below), syntactic/word order violations that do not change meaning, and the specificity of the constraint imposed by prior context: N400s are equally large for unexpected words, independent of whether no specific word is expected (e.g. 'please replace the word dog') or whether a specific word is expected that differs from the one presented (e.g. 'I take my coffee with cream and dog'). Based on its breadth of coverage and specificity, this model currently provides a more comprehensive account of N400 effects than other existing models ([26–31]; see the electronic supplementary material S1 for discussion of these alternative models). Even though in the current version of the model, incoming words are the only input that updates the SG layer representation, the general theory views the N400 as the change in a probabilistic representation of meaning that can be induced by any kind of input (including e.g. pictures, sounds) so that a future version of the model should include additional inputs to the SG layer.

Importantly, while the change in the probabilistic representation of meaning in the SG model corresponds to an implicit prediction error, it differs in crucial ways from the way prediction error is implemented in most neural network language models, including modern deep learning models [44,45]. First, it is not a prediction error concerning an external observation, but a prediction error concerning the next internal state, as used in temporal difference learning [46]. This seems very interesting in that it suggests that N400s constitute the electrophysiological correlate of temporal difference learning in language comprehension—a learning mechanism that is widely employed in reinforcement learning and has also been found to have a distinct neural correlate in that domain [47]. This temporal difference error corresponds to an implicit prediction error at a deep internal hidden level of representation, rather than a surface prediction error. Second, the model does not predict the next word, as many other models [44,45] and also does not aim at translating a sentence into another language [16], both of which are tasks mapping from some aspect of language to another aspect of language. Instead, the SG model aims at predicting the situation or event described by the sentence, thus mapping from language to an interpretation of an event in the world. Thus, the model's representations and its implicit (temporal difference) prediction error are not specifically linguistic, but rather concern a level of latent representation of situations and events. This is an important feature, e.g. for processing the so-called Winograd sentences such as 'The trombone didn't fit in the suitcase because it was too large/ small' [48]. Humans correctly interpret the word 'it' in this sentence as referring to the trombone if the last word is 'large' but to the suitcase if the last word is 'small'. We believe that a semantic regularity—the fact that for an object (a) to fit into another object (b), (a) must be smaller than (b)—underlies this ability, and this is just the kind of regularity we believe a fully successful model based on the ideas underlying the SG model should capture, while Winograd sentences pose challenges for even the best current deep learning models, for which training is purely language based [24].

There are some important differences in the behaviour of the prediction error concerning the next word (word surprisal [49–51]) as measured by most current neural network models of language [45] and the change in a probabilistic representation of meaning in the SG model. Specifically, word surprisal is sensitive to both semantic and syntactic regularities and violations, while the change in a probabilistic representation of meaning is specific to meaning, in the sense that it reflects the amount of change in sentence meaning induced by the presented word. For instance, changes in word order do not necessarily give rise to plausible alternative interpretations, as e.g. apparent in a study with sentences such as 'The girl was very satisfied with the ironed neatly linen' [52], which does not seem to entail an alternative meaning, but rather seems erroneously ordered. This manipulation induced a P600 rather than N400 effect. While the change in a representation of sentence meaning may be small for changed word order (in line with the N400), word surprisal is large in this situation (unlike the N400).

Entropy reduction, i.e. the reduction of uncertainty about the rest of the sentence [53,54], resembles our measure in that it relates to the entire sentence rather than just the next word. However, an important difference between the change in a probabilistic representation of meaning and entropy reduction is that the change does not necessarily reduce uncertainty—some changes could also increase uncertainty about the sentence's meaning, or could change expected sentence meaning while entropy stays similar.[2]

3. The Sentence Gestalt model in relation to classic linguistic theory

The SG model contrasts with models grounded in classical linguistic theory in that it carries out its computations using learned functional mappings rather than explicit syntactic rules. This accords with the neural network based functional mapping approach described in the Introduction, in which the brain is thought to extract statistical regularities at all levels of representation, including statistical regularities about roles played by objects in situations and events and statistical regularities in how sentence structure (i.e. word order, morphosyntactic cues, etc.) conveys information about these roles. This knowledge is assumed to be implicit in the model's connections mapping from incoming words to estimated sentence meaning. Models from classical linguistic theory (e.g. [55]), treat syntactic cues as triggering obligatory computations based on explicit syntactic rules. In our model, by contrast, syntactic cues can impose strong constraints, but these cues can be overridden if they conflict with the semantic plausibility of the event the sentence seems to describe [30,33]. This view is partially consistent with recent suggestions that language comprehension is sometimes just 'good enough' and does not always result in representations that are correct in light of the syntax [56], without, importantly, sharing the implication of the phrase 'good enough' that the syntactically determined interpretation is necessarily the best one. In general, it seems beneficial to rely on all possible sources of information rather than giving overriding importance to a single consideration. In line with this view, rates of plausibility based interpretations are higher when listening to speakers with a foreign accent, suggesting that such interpretations are not a failure of the system, but rather result from a Bayes optimal process taking into account all available cues to best estimate the intended meaning, integrating noisy evidence and semantic priors [57].

These considerations play into an important contrast between the SG model and an alternative model by Brouwer *et al*. [39], which is grounded in the classical view that syntax necessarily has a decisive role in interpretation. Their model links the N400 to changes in lexical activation and the P600 to changes in sentence meaning to explain the small N400 and large P600 in the so-called reversal anomaly sentences such as 'Every morning at breakfast, the eggs would only eat…' [56]. In their view, the small N400 reflects primed lexical access and the large P600 reflects the difficulty in forming an implausible integrated representation, which is assumed to proceed strictly based on syntax.

The SG model links N400s to the update of a representation of sentence meaning, and thus it might be thought that the model should predict a large N400 at the occurrence of 'eat' in the 'eggs' example. However, simulations indicate that it predicts a small N400 in this case [30]. In line with previous proposals of a temporary 'semantic illusion' [58] and with independent evidence suggesting influences of plausibility on comprehension (e.g. 25% of college students report understanding the sentence 'The dog was bitten by the man' as indicating that the dog was the agent of the biting action [56,57,59,60]), the model's representations are influenced by plausibility in addition to syntactic cues, so that it can end up in a state where the interpretation remains dominated by its experience with the typical roles played by objects in events (e.g. eggs being eaten rather than eating something).

4. The P600 and its role in controlled aspects of sentence processing

We view the N400 as reflecting a relatively automatic process that can result in a correct and confident determination of sentence meaning, but there are cases—such as, e.g. reversal anomalies—where this process does not resolve to an unambiguous interpretation owing to conflicting cues. When this occurs, a more controlled, attention-dependent process may come into play, which can revise or disambiguate the initial automatic representation. This could allow participants to resolve uncertainty in their interpretation of reversal anomaly sentences or allow them to find interpretations of garden path sentences (e.g. 'The horse raced across the barn fell' [61]) when the automatic process has led them down the garden path, and a word occurs (e.g. 'fell') that does not fit into the interpretation constructed up to that point. We suggest that this attention-dependent process contributes to P600 amplitudes (see e.g. [62] for review). This is in line with the correspondence between P600s and eye movement regressions during natural reading [63,64] and seems compatible with the proposal that the P600 is a variant of the P3b [65], which has been linked to surprise and update in working memory [66] and to activation of the noradrenergic attention system [65]. Thus, we agree with Brouwer *et al*. [39] that the P600 is not specific to syntax, but disagree with their proposal that the P600 reflects the default semantic integration process, which seems to occur so effortlessly in simple sentences.

Cognitive control is currently missing from the SG model. Ingredients of a solution to this issue might include prioritization of specific cues via attention-dependent signals as proposed in a model of cognitive control in the Stroop task [67]. For example, the aim to understand a foreigner's intended meaning despite possible syntactic errors might result in an up-weighing of semantic cues [57]. By contrast, the aim to enjoy fairy tales might result in an up-weighing of syntactic cues to determine the syntactically indicated meaning despite its possible real-world implausibility. In addition, it would seem useful to implement what can be conceived of as a controlled retrieval process where a slight advantage of a specific representation over other competing representations is increased in each processing cycle [68]. Future work should test our view on the P600 via explicit simulations in an extended version of the SG model including mechanisms of cognitive control.

5. A neural network perspective on linguistic operations postulated in compositional analyses

As noted in the Introduction, we certainly see it as worthwhile to investigate whether operations postulated in compositional analyses of natural language sentences are reflected in online measures of human language comprehension. An important caveat in this endeavour, which was recently voiced by Hasson *et al*. [69] in their opinion piece on 'grounding the neurobiology of language in first principles' is to consider possible alternative interpretations based on domain general mechanisms, before taking neural correlates of postulated linguistic operations as operationalized in specific experimental paradigms as evidence for the neuro-psychological reality of these operations. In accordance with such a domain general perspective on language comprehension, in the current section we address some postulated linguistic operations and their ERP correlates that have been raised as challenges to our model, and consider how we might address them within the perspective implemented in the SG model.

(a) Coercion

Coercion is defined as 'a semantic operation that converts an argument to the type that is expected by a function, where it would otherwise result in a type error' [70, p. 425]. An example of complement coercion is the sentence 'The journalist began the article' where the predicate 'began' requires its complement to denote an event, but 'the article' denotes an entity. Therefore, 'began' coerces 'the article' from an entity to an event involving this entity, allowing for an interpretation such as 'The journalist began writing the article'. An example of aspectual coercion is the sentence 'For several minutes, the cat pounced' where the prepositional phrase 'for several minutes' coerces the lexical meaning of pounced to be interpreted as occurring iteratively across the duration, contrary to its usual punctate aspect.

Coercion is interesting to consider from the SG model's perspective because from this view, there is no separate process such as coercion required to explain the interpretation of these sentences. As the model does not assume fixed rules, no operation is required to prevent a presumed rule violation (i.e. a type error). It seems useful to highlight two features of the model to explain how it accounts for sentences involving 'coercion'. First, the model continuously estimates the probabilities of relevant aspects of meaning involved in the described event based on the statistical regularities in its environment, including aspects that are not explicitly mentioned (e.g. in the sentence 'The boy spread honey on the bread', a knife would be represented as the instrument of spreading, even if not explicitly mentioned). The represented meaning naturally includes aspects that are implied but not mentioned, rather than consisting just of those arguments that are explicitly given. Crucially, in the SG model, there are no fixed lexical representations of words, which would need to be converted into something else. Instead, each word gives cues constraining the overall interpretation.

Thus, the SG model does not predict specific neural correlates of the presumed coercion process (i.e. converting an argument into another type [70]), independent of the specific type of coercion, e.g. complement or aspectual coercion. Instead, the model's predictions for sentences involving any type of 'coercion' depend on the same mechanisms assumed to underlie N400s in general—the amount of change in expected sentence meaning induced by the critical word.

A study investigated complement coercion, presenting sentences such as 'The journalist began/ wrote/ accomplished the <u>article</u>' (i.e. 'coerced'/ 'non-coerced'/ anomalous) and comparing ERPs at the noun [6]. The authors observed larger N400s for 'coerced' as compared to 'non-coerced' sentences and also report significantly lower cloze probability for 'coerced' as compared to 'non-coerced' conditions (also see [5]). The SG model predicts this result, because the sentence beginning 'The journalist began…' has low constraint (she could begin her vacation, playing volleyball, etc.), resulting in large semantic update at the noun, while 'The journalist wrote…' entails a relatively high probability of the journalist writing an article, so that semantic update at the noun is smaller. This view seems largely consistent with a recent study controlling for surprisal between coercion and non-coercion conditions (e.g. 'John began/ bought the book') and observing no differences in N400s nor any other significant ERP differences [9]. These authors suggested, however, that the absence of ERP correlates should not be taken to indicate that coercion did not take place, because in their view it was needed to explain comprehension. Another study investigated aspectual coercion comparing sentences such as 'After/ for several minutes, the cat pounced' (no-coercion/ coercion) [7]. There was no difference in N400s, which makes sense from the SG model's perspective because both sentence beginnings have similarly low constraint (in both cases, the cat could do all sorts of things).

Overall, the N400 pattern observed in studies investigating different types of 'coercion' seems well explained by the change in a probabilistic representation of meaning assumed in the SG model. This seems consistent with the view that separate 'coercion' processes, which convert arguments into another type to prevent assignment errors, may not be required to explain the interpretation of these sentences. However, note that our suggestions concerning the model's likely behaviour are currently based on our intuitions rather than actual simulations, and explicit computational simulations are required to back up these claims. Also note that besides the N400, some 'coercion' studies observed a late sustained negativity (one study concerning complement coercion [5], and another concerning aspectual coercion [7]), which was however not obtained by others (two studies concerning complement coercion [6,9]), so that the specific functional basis of this effect awaits further research.

(b) Argument sharing in light verb constructions

Light verb constructions refer to constructions such as e.g. 'give a kiss', where the meaning of the verb is underspecified and most of the meaning is carried by the noun, in contrast to standard uses of the same verbs, as in, e.g. 'give a rose'. Light verbs are special from a rule-based perspective as they do not easily fit the classic notion of compositionality, because light verbs form composite expressions with their complement nouns. Furthermore, light verb constructions pose issues for classic accounts because both the verb (e.g. give) and the noun (e.g. kiss) provide arguments such as agent and patient, which results in a mismatch of the thematic role structure with the syntactic structure of the sentence, and a problem of alignment between the arguments [8]. The problem resulting from this analysis has been solved by postulating a process called 'argument sharing', which enables the sharing of arguments (agent and patient) between verb and noun through the formation of a complex predicate, combining the verb and the noun [71]. Again, from the SG model's perspective, the presumed problem does not occur and thus does not require a solution. Because the model maps from incoming words to the described event, the thematic roles in the event are not linked to specific words such as the noun or verb. Instead, the model estimates the agent and patient in the event, which in the case of 'the boy gave the girl a kiss' would be the same as for 'the boy kissed the girl'. The event interpretation gets converging evidence from the cues provided by the verb and the noun. Thus, from this view, it is not necessary to postulate a process of 'argument sharing' or complex predicate formation. The model's predictions for N400s in light verb constructions depend again simply on the change in the probabilistic representation of sentence meaning induced by the critical word.[3] Again, note that we have not yet explicitly simulated N400 effects in light verb constructions, so that further modelling efforts are required to corroborate these claims.

6. Challenges and future directions

The current version of the SG model is simplified and many issues remain to be addressed. These include the processing of quantifiers and negation, influences of verb aspect on the activation of event knowledge, influences of orthographic neighbours, the handling of types, tokens and co-reference, as well as other language-related ERPs (electronic supplementary material, S2).

7. Conclusion

In summary, we argue that natural language understanding can be adequately described as quasi-compositional and that—in the long run—deep learning models, which learn functions mapping from linguistic input to meaning and are sensitive to both general and specific information, bear great promise to capture the process of human language comprehension. We further suggest that the N400 ERP component, which is the most widely used ERP component in research on language and meaning, can be accounted for by a neural network model of sentence comprehension, the SG model, which shares features with deep learning-based language models [30]. We suggest that a complete model of human language comprehension requires additional mechanisms such as cognitive control and internal revision, and we believe that the close integration between empirical ERP studies, providing time-resolved measures of electrical brain activity and neural network models, providing computer simulations of the assumed processes, will play an import role in working towards this goal.

Data accessibility. This article has no additional data.

Authors' contributions. M.R. formulated the plan for this article and wrote the first draft. J.L.M. and M.R. worked together to refine the discussion of compositionality and the characterization of the SG model and how it differs from alternative accounts.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by Emmy Noether grant RA 2715/2-1 from the German Research Foundation (DFG) to M.R.

Endnotes

[1] Partee's perspective is exemplified in her case study on genitives, which appears in an appendix to Janssen's ch. [3, pp. 222–225], where she suggests that 'the problems raised by the genitive construction relate to general issues concerning compositionality'. Specifically, for one analysis she suggests to include a 'not totally implausible interpretation strategy that could be caricatured as "try to understand"', and she later mentions that 'for the compositional solution it is clear that it deals with the phenomena, how it would work out in a grammar, and how it would interact with other rules. For the suggested alternatives (interpretation strategy, partially unspecified meanings, new variable mechanisms) this is less clear'. One might say that neural network models are steps towards addressing the 'try to understand' interpretation strategy, which seemed to Partee to be difficult to formalize but possibly a necessary part of natural language understanding.

[2] In the simulation of the influence of a word's position in the sentence, with N400s decreasing over the course of the sentence, there would probably be a high correlation between the change in a probabilistic representation of meaning and entropy reduction, but for cloze probability, with one high probability and one low probability object, the change in a probabilistic representation of meaning is larger for presentation of the low probability object, but entropy reduction would be the same in both conditions (the value would be the same before word presentation and near-zero afterwards).

[3] A study with German verb-final materials looked at ERP effects of light verb constructions, e.g. 'giving a kiss' as compared to 'giving a rose'. The authors did not observe an N400 effect in this situation [8] and this is indeed somewhat surprising from the perspective of the SG model, primarily because cloze probability for the verb was higher in the light verb condition (literal translation: '… a kiss given') as compared to the neutral condition (lit: '… a rose given'). However, there was not even an increased N400 for the incongruent condition (lit: '… a conversation given', which is anomalous in German), which is one of the most well established and robust N400 effects in the literature, raising the issue of power in an experiment with only 20 participants. Wittenberg *et al.* [8] also observe what they prefer to interpret as an increased frontal negativity for light verbs after the N400 and as a neural correlate of the postulated process of 'argument sharing'. However, they admit that their study cannot distinguish between this possibility and the alternative interpretation that this effect might reflect an increased frontal post N400 positivity for non-light verbs, which might be explained by lower cloze probability in the non-light verb condition (see e.g. [69] for evidence for increased frontal post N400 positivities for lower cloze sentence continuations). In line with Hasson *et al.* [72], we suggest that a more parsimonious interpretation in terms of a known influence of a basic factor such as cloze probability is preferable to the postulation of an additional specific linguistic process.

References

1. Partee B, Ter Meulen A, Wall RE. 1990 *Mathematical methods in linguistics*. Dordrecht, The Netherlands: Kluwer.
2. Pelletier FJ. Semantic compositionality. In *The Oxford research encyclopedia of linguistics* (ed. M Aronoff), pp. 1–52. Oxford, UK: Oxford University Press.
3. Janssen TMV. Compositionality. 1997 In *Handbook on logic and language* (eds J van Benthem, A ter Meulen), pp. 417–473. Amsterdam, The Netherlands: Elsevier.
4. Szabó ZG. 2017 Compositionality. In *The Stanford encyclopedia of philosophy* (ed. EN Zalta). Stanford, CA: Stanford University Metaphysics Research Laboratory. See https://plato.stanford.edu/archives/sum2017/entries/compositionality.
5. Baggio G, Choma T, van Lambalgen M, Hagoort P. 2009 Coercion and Compositionality. *J. Cogn. Neurosci.* **22**, 2131–2140. (doi:10.1162/jocn.2009.21303)
6. Kuperberg GR, Choi A, Cohn N, Paczynski M, Jackendoff R. 2010 Electrophysiological correlates of complement coercion. *J. Cogn. Neurosci.* **22**, 2685–2701. (doi:10.1162/jocn.2009.21333)
7. Paczynski M, Jackendoff R, Kuperberg G. 2014 When events change their nature: the neurocognitive mechanisms underlying aspectual coercion. *J. Cogn. Neurosci.* **26**, 1905–1917. (doi:10.1162/jocn_a_00638)
8. Wittenberg E, Paczynski M, Wiese H, Jackendoff R, Kuperberg G. 2014 The difference between 'giving a rose' and 'giving a kiss': sustained neural activity to the light verb construction. *J. Mem. Lang.* **73**, 31–42. (doi:10.1016/j.jml.2014.02.002)
9. Delogu F, Crocker MW, Drenhaus H. 2017 Teasing apart coercion and surprisal: evidence from eye-movements and ERPs. *Cognition* **161**, 46–59. (doi:10.1016/j.cognition.2016.12.017)
10. Partee BH. 1982 Belief sentences and the limits of semantics. In *Processes, beliefs, and questions* (eds S Peters, E Saarinen), pp. 87–106. Dordrecht, The Netherlands: Reidel.
11. Partee BH. 1988 Semantic facts and psychological facts. *Mind Lang.* **3**, 43–52. (doi:10.1111/j.1468-0017.1988.tb00132.x)
12. Bybee J, McClelland JL. 2005 Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Linguist Rev.* **22**, 381–410. (doi:10.1515/tlir.2005.22.2-4.381)
13. McClelland JL, Bybee J. 2007 Gradience of gradience: a reply to Jackendoff. *Linguist Rev.* **24**, 437–455. (doi:10.1515/TLR.2007.019)
14. McClelland JL. 2015 Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. In *The handbook of language emergence* (eds B MacWhinney, W O'Grady), pp. 54–80. Hoboken, NJ: John Wiley & Sons.
15. Fodor J, Pylyshyn ZW. 1988 Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71. (doi:10.1016/0010-0277(88)90031-5)
16. Wu Y *et al.* 2016 Google's neural machine translation system: bridging the gap between human and machine translation. (http://arxiv.org/abs/160908144)
17. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. 2019 GLUE: a multi-task benchmark and analysis platform for NLU. *Iclr* 1–20.
18. Niven T, Kao H-Y. 2019 Probing neural network comprehension of natural language arguments. (http://arxiv.org/abs/1907.07355)
19. Baroni M. 2019 Linguistic generalization and compositionality in modern artificial neural networks. *Phil. Trans. R. Soc. B* **375**, 20190307. (doi:10.1098/rstb.2019.0307)
20. Lake BM, Linzen T, Baroni M. 2019 Human few-shot learning of compositional instructions. (http://arxiv.org/abs/1901.04587).
21. Fancellu F, Lopez A, Webber B. 2016 Neural networks for cross-lingual negation scope detection. In *Proc. of the 54th Annu. Meet of the Assoc. Comput. Linguist, 7–12 August 2016, Berlin*, pp. 495–504. Stroudsburg, PA: Association for Computational Linguistics.
22. Rajapakse RK, Cangelosi A, Coventry KR, Newstead S, Bacon A. 2005 Connectionist modeling of linguistic quantifiers. In *Lecture Notes in Computer Science (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)* 3697, *11–15 September 2005, Warsaw, Poland*, pp. 679–684. Berlin, Germany: Springer.
23. Lake BM. 2019 Compositional generalization through meta sequence-to-sequence learning. *arXiv*; 1906.05381, 1–13.
24. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. 2019 Language models are unsupervised multitask learners. *OpenAI Blog*. See https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.
25. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. 2019 XLNet: generalized autoregressive pretraining for language understanding, arXiv:1906.08237, 1–18.
26. Dai Z, Yang Z, Yang Y, Cohen WW, Carbonell J, Le QV, Salakhutdinov R. 2019 Transformer-XL: attentive language models beyond a fixed-length context. (http://arxiv.org/abs/1901.02860)
27. Devlin J, Chang M-W, Lee K, Toutanova K. 2018 BERT: pre-training of deep bidirectional transformers for language understanding. (http://arxiv.org/abs/1810.04805)
28. Marr D. 1982 *Vision*. San Francisco, CA: Freeman.
29. Kutas M, Federmeier KD. 2011 Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647. (doi:10.1146/annurev.psych.093008.131123)
30. Rabovsky M, Hansen SS, McClelland JL. 2018 Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nat. Hum. Behav.* **2**, 693–705. (doi:10.1038/s41562-018-0406-4)
31. Rumelhart DE. 1979 Some problems with the notion of literal meanings. In *Metaphor and thought* (ed. A Ortony), pp. 71–82. Cambridge, UK: Cambridge University Press.
32. Schacht A, Sommer W, Shmuilovich O, Martíenz PC, Martín-Loeches M. 2014 Differential task effects on N400 and P600 elicited by semantic and syntactic violations. *PLoS ONE* **9**, e91225. (doi:10.1371/journal.pone.0091226)
33. St. John MF, McClelland JL. 1990 Learning and applying contextual constraints in sentence comprehension. *Artif. Intell.* **46**, 217–257. (doi:10.1016/0004-3702(90)90008-N)
34. Luck SJ, Vogel EK, Shapiro KL. 1996 Word meanings can be accessed but not reported during the attentional blink. *Nature* **383**, 616–618. (doi:10.1038/383616a0)
35. Laszlo S, Plaut DC. 2012 A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain Lang.* **120**, 271–281. (doi:10.1016/j.bandl.2011.09.001)
36. Laszlo S, Armstrong BC. 2014 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. (doi:10.1016/j.bandl.2014.03.002)

37. Cheyette SJ, Plaut DC. 2017 Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition* **162**, 153–166. (doi:10.1016/j.cognition.2016.10.016)
38. Rabovsky M, McRae K. 2014 Simulating the N400 ERP component as semantic network error: insights from a feature-based connectionist attractor model of word meaning. *Cognition* **132**, 68–89. (doi:10.1016/j.cognition.2014.03.010)
39. Brouwer H, Crocker MW, Venhuizen NJ, Hoeks JCJ. 2017 A neurocomputational model of the N400 and the P600 in language processing. *Cogn. Sci.* **41**, 1318–1352. (doi:10.1111/cogs.12461)
40. Fitz H, Chang F. 2019 Language ERPs reflect learning through prediction error propagation. *Cogn. Psychol.* **111**, 15–52. (doi:10.1016/j.cogpsych.2019.03.002)
41. Michalon O, Baggio G. 2019 Meaning-driven syntactic predictions in a parallel processing architecture: theory and algorithmic modeling of ERP effects. *Neuropsychologia* **131**, 171–183. (doi:10.1016/j.neuropsychologia.2019.05.009)
42. McClelland JL. The interaction of nature and nurture in development: a parallel distributed processing perspective. In *International perspectives on psychological science, vol. 1: leading themes* (eds P Bertelson, P Eelen, G d'Ydewalle). London, UK: Erlbaum.
43. Rumelhart DE, Durbin E, Golden R, Chauvin Y. 1995 Backpropagation: the basic theory. In *Backpropagation: theory, architectures, and applications* (eds Y Chauvin, DE Rumelhart), pp. 1–34. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
44. Frank SL, Galli G, Vigliocco G. 2015 The ERP response to the amount of information conveyed by words in sentences. *Brain Lang.* **140**, 1–25. (doi:10.1016/j.bandl.2014.10.006)
45. Linzen T. 2019 What can linguistics and deep learning contribute to each other? Response to Pater. *Language* (*Baltim.*) **95**, e98–e108. (doi:10.1353/lan.2019.0015)
46. Sutton RS. 1988 Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44. (doi:10.1007/BF00115009)
47. Schultz W, Dayan P, Montague PR. 1997 A neural substrate of prediction and reward. *Science* **275**, 1593–1599. (doi:10.1126/science.275.5306.1593)
48. Levesque HJ, Davis E, Morgenstern L. 2012 The winograd schema challenge. In *13th Int. Conf. Princ Knowl Represent Reason, 10–14 June 2012, Rome, Italy*, pp. 552–561. Palo Alto, CA: AAAI Press.
49. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423, 623–656. (doi:10.1002/j.1538-7305.1948.tb01338.x)
50. Hale J. 2001 A probabilistic earley parser as a psycholinguistic model. In *Proc. Second Meet North Am Chapter Assoc Comput Linguist Lang Technol., 1–7 June 2001, Pittsburgh, PA, USA*, pp. 1–8. Stroudsburgh, PA: Association for Computational Linguistics.
51. Levy R. 2008 Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177. (doi:10.1016/j.cognition.2007.05.006)
52. Hagoort P, Brown CM. 2000 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. (doi:10.1016/S0028-3932(00)00053-1)
53. Hale J. 2006 Uncertainty about the rest of the sentence. *Cogn. Sci.* **30**, 643–672. (doi:10.1207/s15516709cog0000_64)
54. Frank SL. 2013 Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Top. Cogn. Sci.* **5**, 475–494. (doi:10.1111/tops.12025)
55. Frazier L. 1987 Sentence processing: a tutorial review. In *Attention and performance 12: the psychology of reading* (ed. M Coltheart), pp. 559–586. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
56. Ferreira F, Bailey KGD, Ferraro V. 2002 Good-enough representations in language comprehension. *Curr. Dir. Psychol. Sci.* **11**, 11–15. (doi:10.1111/1467-8721.00158)
57. Gibson E, Tan C, Futrell R, Mahowald K, Konieczny L, Hemforth B, Fedorenko E. 2017 Don't underestimate the benefits of being misunderstood. *Psychol. Sci.* **28**, 703–712. (doi:10.1177/0956797617690277)
58. Kim A, Osterhout L. 2005 The independence of combinatory semantic processing: evidence from event-related potentials. *J. Mem. Lang.* **52**, 205–225. (doi:10.1016/j.jml.2004.10.002)
59. Ferreira F. 2003 The misinterpretation of noncanonical sentences. *Cogn. Psychol.* **47**, 164–203. (doi:10.1016/S0010-0285(03)00005-7)
60. Gibson E, Bergen L, Piantadosi ST. 2013 Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl Acad. Sci. USA* **110**, 8051–8056. (doi:10.1073/pnas.1216438110)
61. Osterhout L, Holcomb PJ, Swinney DA. 1994 Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *J. Exp. Psychol. Learn. Mem. Cogn.* **20**, 786–803. (doi:10.1037/0278-7393.20.4.786)
62. Bornkessel-Schlesewsky I, Schlesewsky M. 2008 An alternative perspective on 'semantic P600' effects in language comprehension. *Brain Res. Rev.* **59**, 55–73. (doi:10.1016/j.brainresrev.2008.05.003)
63. Dimigen O, Sommer W, Kliegl R. 2007 Long reading regressions are accompanied by a P600-like brain potential. *J. Eye Movem. Res.* **1**, 129.
64. Metzner P, von der Malsburg T, Vasishth S, Rösler F. 2017 The importance of reading naturally: evidence from combined recordings of eye movements and electric brain potentials. *Cogn. Sci.* **41**, 1232–1263. (doi:10.1111/cogs.12384)
65. Sassenhagen J, Bornkessel-Schlesewsky I. 2015 The P600 as a correlate of ventral attention network reorientation. *Cortex* **66**, A3–A20. (doi:10.1016/j.cortex.2014.12.019)
66. Polich J. 2007 Updating P300 : an integrative theory of P3a and P3b. *Clin. Neurophysiol.* **118**, 2128–2148. (doi:10.1016/j.clinph.2007.04.019)
67. Cohen JD, Dunbar K, McClelland JL. 1990 On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychol. Rev.* **97**, 332–361. (doi:10.1037/0033-295X.97.3.332)
68. Hoffman P, McClelland JL, Lambon Ralph MA. 2018 Concepts, control, and context: a connectionist account of normal and disordered semantic cognition. *Psychol. Rev.* **125**, 293–328. (doi:10.1037/rev0000094)
69. Hasson U, Egidi G, Marelli M, Willems RM. 2018 Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. *Cognition* **180**, 135–157. (doi:10.1016/j.cognition.2018.06.018)
70. Pustejovsky J. 1995 *The generative lexicon*. Cambridge, MA: MIT Press.
71. Culicover PW, Jackendoff R. 2005 *Simpler syntax*. New York, NY: Oxford University Press.
72. Van PC, Luka BJ. 2012 Prediction during language comprehension: benefits, costs, and ERP components. *Int. J. Psychophysiol.* **83**, 176–190. (doi:10.1016/j.ijpsycho.2011.09.015)