PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Introduction

Cite this article: Martin AE, Baggio G. 2019 Modelling meaning composition from formalism to mechanism. *Phil. Trans. R. Soc. B* **375**: 20190298.
http://dx.doi.org/10.1098/rstb.2019.0298

Accepted: 28 October 2019

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

Subject Areas:
cognition, neuroscience, behaviour

Keywords:
semantics, language, compositionality, mechanistic models, cognition, neuroscience

Author for correspondence:
Giosuè Baggio
e-mail: giosue.baggio@ntnu.no

THE ROYAL SOCIETY PUBLISHING

Modelling meaning composition from formalism to mechanism

Andrea E. Martin[1,2] and Giosuè Baggio[3]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]Donders Centre for Cognitive Neuroimaging, Radboud University, Nijmegen, The Netherlands
[3]Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway

AEM, 0000-0002-3395-7234; GB, 0000-0001-5086-0365

Human thought and language have extraordinary expressive power because meaningful parts can be assembled into more complex semantic structures. This partly underlies our ability to compose meanings into endlessly novel configurations, and sets us apart from other species and current computing devices. Crucially, human behaviour, including language use and linguistic data, indicates that composing parts into complex structures does not threaten the existence of constituent parts as independent units in the system: parts and wholes exist simultaneously yet independently from one another in the mind and brain. This independence is evident in human behaviour, but it seems at odds with what is known about the brain's exquisite sensitivity to statistical patterns: everyday language use is productive and expressive precisely because it can go beyond statistical regularities. Formal theories in philosophy and linguistics explain this fact by assuming that language and thought are *compositional*: systems of representations that separate a variable (or *role*) from its values (*fillers*), such that the meaning of a complex expression is a function of the values assigned to the variables. The debate on whether and how compositional systems could be implemented in minds, brains and machines remains vigorous. However, it has not yet resulted in mechanistic models of semantic composition: how, then, are the constituents of thoughts and sentences put and held together? We review and discuss current efforts at understanding this problem, and we chart possible routes for future research.

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

1. Meaning composition

Natural language and other symbolic systems, such as logic and mathematics, are *combinatorial* and *compositional*. A system where simpler symbols can be put together into more complex symbols in systematic ways is combinatorial: in natural language, morphemes combine into words, and words into phrases and sentences. A system where combining symbols also results in combining their meanings, again in systematic ways, is compositional: given a phrase or sentence, it is very often possible to assign to it a meaning that is a function of the meanings of the parts (e.g. the constituent words) and of the structure of the whole expression [1,2]. Meaning composition is remarkable among human mental capacities and behaviours, because it does not appear to be adequately accounted for by statistical relationships or by associative processing alone [3–6]. This fact stands in striking contrast with the behaviour of other perception–action and cognitive systems which can be well predicted by traditional or contemporary statistical or associative models. A further vexing conundrum lies in explaining how systems in the brain realize meaning composition within the bounds of neurophysiological computation, given that the human brain is a computational device whose primary remit is to learn from and capitalize upon the statistical structure of its environment.

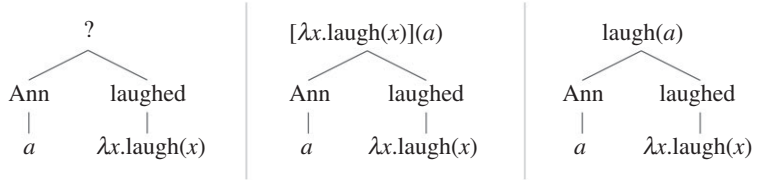

Figure 1. Meaning composition as function application in formal semantics.

Compositionality is an issue of much discussion in linguistics and philosophy (e.g. [7–10]): it is a matter of debate to what extent languages, as opposed to semantic theories, are compositional. But whether compositionality holds strongly or weakly for given formal or natural languages, there is no question that the meanings of complex expressions may be derived in systematic ways from the meanings of the parts. Meanings can be composed, and composition must ultimately boil down to an easily computable or tractable mathematical function [10–12].

The standard approach in formal semantics is to model meaning composition as *function application* [13,14], as illustrated in figure 1 for the sentence 'Ann laughed'. The proper name 'Ann' is represented by a constant (a) and the predicate 'laughed' by a lambda term, i.e. a function that may be applied to arguments of the appropriate type (left panel). How is the meaning of 'Ann laughed' composed? The function 'laughed' is applied to the argument 'Ann' (middle panel, top), and in the resulting expression (right panel, top) the bound variable (x), in the body of the lambda term, is replaced with the argument expression (a). The idea here is that meaning composition always consists in the 'saturation of an unsaturated meaning component', an idea also known as 'Frege's conjecture' [15, p. 3]. Function application in the lambda calculus gives a precise account of this 'saturation' process. This approach has been hugely influential in linguistics, philosophy, logic and computer science. It has also been strikingly successful in showing how in principle a seemingly complex aspect of human linguistic competence can be captured by a single universal operation in a system of formal logic. To the extent that expressions in natural language can be modelled as predicates and arguments, a variable in a lambda term can be replaced by *any* argument of the right type. And indeed human language is *productive*: we could say of *anyone* that they laughed, because our semantic representation of laughing is *independent* of the individuals of whom laughing is predicated.

Yet, the power and expressivity of the traditional formal analysis of meaning composition come at a high cost: rampant idealization. It is not clear just *how semantic* composition is in the standard lambda calculus model, for two main reasons. First, recursive function application produces formulae in predicate logic (or some high-order intensional language), which should be *interpreted* in a given domain or reference structure: is it function application as such or interpretation that yields meaning? This question disappears only in a theory in which composition and interpretation closely mirror each other, i.e. where a strong version of compositionality holds. Second, it is not clear how lambda terms specify the meanings of constituent expressions. Consider the meaning of 'laughed', according to figure 1. What the lambda term does here is merely lift the relevant expression from the object language to the metalanguage. A lexical meaning is just a placeholder for a typed entity [16,17], and meaning composition therefore reduces to logico-syntactic composition of placeholders for typed entities.

Recent trends in linguistics, psychology, neuroscience and computer science suggest that different formalisms may be used to model lexical meaning, with important consequences for theories of semantic composition. In all of these formalisms, lexical meanings are richer data structures than lambda terms in formal semantics. The idea of breaking down lexical meanings into 'atoms', or in any case simpler constituents, goes back to decompositional approaches in lexical semantics [18,19], variously elaborated and refined in subsequent work on conceptual semantics [20], the generative lexicon [21] and beyond. This idea has gained much traction in recent years with the development of vector-based analyses of meaning in distributional semantics and related approaches [22–24] and with renewed interest in artificial neural networks that can be trained to represent lexical meanings as vectors. In spite of deep differences between these approaches, they share a common implication: if lexical meanings are rich, internally elaborated data structures (conceptual structures, qualia or event or argument structures, distributional vectors, etc.), meaning composition may be difficult or impossible to describe in terms of function application or the 'saturation of an unsaturated meaning component'. This raises several fundamental questions about the nature and scope of composition, i.e. whether it is a simple or a complex function, one or many operations, autonomous or fully reliant on syntax, etc.

From Frege to deep learning, the history of research on meaning composition reveals a succession of different theoretical approaches, using various formal technologies, from mathematical logic to linear algebra and statistics. Most of these formalisms have strived to meet two general requirements:

(1) that lexical meanings, regardless of how they are formalized, are rich semantic representations that can account for all the complexities and nuances of human lexical competence.
(2) that meaning composition, regardless of how it is formalized, upholds the independence of predicates and arguments—variables and values, roles and fillers, etc.—as reflected in human linguistic competence.

Different research programmes in the cognitive sciences have emphasized these requirements differently, and have tended to use either of them as a criterion of explanatory success. In generative grammar and formal semantics, but not in other areas of linguistics, (2) has been given priority over (1). In computer science and artificial intelligence (AI) research, but not in psychology or neuroscience, (1) has taken precedence over (2). Yet, a complete, empirically adequate theory of meaning must explain both aspects of human semantic competence [5,25]. This goal is finally visible on the horizon of cognitive science.

This theme issue addresses meaning composition against the background of requirements (1) and (2), and from the vantage point of recent developments in model-based cognitive science. Given the extraordinary range of formalisms that

have been proposed for modelling syntactic and semantic composition in languages, a new need arises for empirically and computationally grounded research on the *algorithmic and neural bases of composition*: current (known) formalisms can guide inquiry into (unknown) mechanisms, and novel results on the algorithmic and neural implementation of the relevant operations can help select among alternative formal theories in logic, syntax and semantics. The long-term goal is an integrative framework, in which theories of meaning composition are connected seamlessly across levels of analysis [26]. The 15 contributions presented in this theme issue are intended to make headway towards this ambitious goal, and to do so in a way that is consistent with what is known about neural computation in the human brain.

2. Mechanistic models

Recent progress in the cognitive and brain sciences suggests that mechanistic models of syntactic and semantic composition are not only possible (i.e. that formal and computational tools exist that allow us to articulate and test such models), but also necessary for further advancement. First, new data analytic techniques have been used in human neuroscience, such as machine learning approaches to decoding the representational structure and contents of brain states. We are now able to probe *neural compositionality*, that is, whether the cortical representations of complex symbols are encoded as functions of the representations of their constituents [27]. Mechanistic models are now needed to guide this endeavour by specifying what functions (additive, multiplicative, etc.) are required by composition. Second, in recent years, progress has been made in characterizing composition formally and in mapping it in brain space and time by means of functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG) or electroencephalography (EEG). However, algorithmic and neurophysiologically grounded models are now necessary in order to begin to link formalism and experimental data.

Mechanistic models are also a requirement for furthering our understanding of complex systems at mesoscopic scales of organization, such as the brain or the systems that constitute it. Some of the earliest attempts at understanding epistemological constraints involved in the study of perception and cognition may be found in Vedic philosophical texts from the Indian subcontinent. The Chandogya Upanishad [28] and Adi Shankaracharya's commentary on it use the metaphor of partial, sensory-deprived exploration of an elephant to highlight the problem:

> perceiving the elephant through touching its different parts, [we] come to have diverse notions regarding it, each one regarding [the elephant] to be like the part that they had touched; and as none of [us] had touched the whole elephant, none had any idea of the elephant as a whole entity [28; §18,29].

This example highlights the inherent methodological challenges of studying a complex system without the constraints of theory and mechanism to guide us: observation may be interpreted only through the lens of the often implicit and unexamined biases of the observer—an issue long at stake in philosophy of science [30]. Newell revisited this problem in his famous 1973 paper, titled 'You can't play 20 questions with nature and win' [31]. He argued that despite, or even because of, rigorous study in putatively disparate fields of psychology or cognitive science, when we study the capacities of the human mind without the infrastructure of formal theory and mechanism, and specifically without the global purview they provide, we are functionally reduced to groping an elephant: we may be getting data from some partial aspects of the phenomenon, without being able to connect those data into a coherent picture, or to formulate an account of the processes that generate those data. This thematic issue espouses the thesis that the purpose of a mechanical model is to accomplish those feats, i.e. to *connect data* and to *explain how data are generated* by the system under scrutiny.

Current formal theories in AI, psychology and neuroscience continue to face dire problems of explanatory adequacy and power, because they are focused on providing either *models of data* (e.g. statistical models) or *models of tasks* obtained via approximation of human or machine behaviour, and not *models of the mechanisms that generate those data* or of system behaviours that have not been (or cannot be) observed in restricted data-collection settings, e.g. a laboratory. This focus, or lack of focus, is exacerbated by the fact that current models are not instantiated within a developing theory of *human capacities*, when in our view, they should be by necessity. Of course contemporary models have made much progress in terms of the tasks they can perform, predict or statistically approximate: this is true in AI, as well as in psychology and neuroscience. But we should emphasize that the question remains as to what, if anything, these models *explain*, and how they achieve their explanatory force, if any, without mechanism. (Deep) neural networks are a glaring example: as the complexity of the task to be accomplished (or problem to be solved) increases, so do the network's structure, size and training parameters; but then it also becomes harder to interpret the network's internal states and go beyond knowledge of the algorithm or rule used to update weights (e.g. back-propagation). This is an instance of a well-known paradox: why pursue modelling, if the model is more complex than the target system? What is the gain of a model that might perform well, but prevents insight into its own workings? And finally, what is a model's value for inquiry, if there is not even the lightest tether to constrain it to solve the problem in the way the mind and brain do?

In order to be explanatory, models require clearly set explananda [30] and must make reasonable simplifications. Models should aim to *represent* observable and partly unobservable events in the brain, and should go beyond predicting or fitting data, capturing the processes that may generate relevant data [32–34]. Bechtel & Abrahamsen [35] define a mechanism as a structure performing some function by virtue of the structure's components and their organization [36]. The mechanisms underlying meaning composition, and human cognition more generally, are both *spatially and temporally organized*: the mechanism's elements have specific spatial properties (i.e. localization in brain or cell-assembly state space, distinctive connectivity, etc.) and temporal properties (i.e. the order, rate, duration of activities, etc.; [37]). The search for the spatio-temporal correlates of meaning composition in the brain, or 'mapping' composition in brain space and time, is a necessary step towards mechanistic models. Further, Kaplan [32] introduces the idea of 'explanatory force' of a model, and he links it to the *causal structure* of the mechanism: 'A model carries explanatory force to the extent that it reveals aspects of the causal structure of a mechanism, and lacks explanatory force to the extent that it fails to describe this structure' [32, pp. 347–348]. In his analysis, if relations between variables in a model account for *how* certain phenomena arise, and if at least some model components correspond to

the 'real-world' mechanism at work, then the model is said to have 'explanatory force'. One diagnostic for this is determining whether through a model one can control or manipulate a phenomenon, not merely predict it. Successful prediction does not entail that the model achieves its results by reproducing the actual mechanism's causal structure. Similarly, the model's ability to successfully predict a phenomenon does not make the model explanatory [32]. Capturing, via a model, the causal structure of the mechanism and its causal roles in producing the phenomena of interest—i.e. developing an account of *why* the variables have the relations that they do, and *how* they generate the observed data—is what endows a model with explanatory force.

We see the challenge for current theories of meaning composition as twofold. On the one hand, the theorist must extract from a broad base of experimental data (M/EEG, fMRI, etc.) information on the spatio-temporal organization of the possible underlying mechanisms. This endeavour is a stepping stone to a full causal analysis of the mechanism at work: here, the aim is to develop an *aetiological explanation* of meaning composition in neural terms, fleshing out the *causal history* of interactions between the relevant brain networks, while also showing that that history achieves the formal requirements espoused by linguistic theory and data. This requires the modeller to go beyond the kind of *descriptive*, *phenomenological* and *constitutive* explanations currently on the market, that is, statements to the effect that such-and-such brain networks or neurophysiological events constitute the relevant mechanisms, and therefore may 'explain' composition (for a discussion, see [38,39]). On the other hand, the task is to develop implementational models *of specific algorithms and computations*, if not necessarily of the actual formalisms used in traditional logical semantics or in contemporary computational linguistics. In other words, *computational explicitness should not be renounced for causal detail*. This is a challenge for all multilevel, integrative efforts in the cognitive sciences (for a discussion of computational mechanisms, see [40,41]; for a key discussion of tractability, see [42]).

The aim of this theme issue is to assess current algorithmic- and neural-level models of composition in brains and machines: are these models mechanistic in the relevant sense? If not, what are the missing ingredients? And do these accounts have sufficient explanatory force to address the human capacity for productive semantic composition, meeting requirements (1) and (2) (§1)? Finally, are these models both causally and computationally explicit?

3. Questions and themes

This theme issue comprises 15 contributions from leading and emerging theorists, modellers and experimentalists in the fields of linguistics, cognitive psychology and neuroscience, AI and computer science. These papers can be clustered together in four groups, each addressing a specific set of questions or themes pertaining to meaning composition in brains and machines.

The first theme is the *neurobiology of meaning composition*: more specifically, the cortical networks supporting composition, or related cognitive processes, in language and beyond. The focus here is on mapping composition-related functions to specific brain regions or networks using neuroimaging methods, such as MEG and fMRI. However, the long-term aim is to use these data types as a springboard for developing explicit and testable models of composition, where information on neural correlates is used to generate hypotheses about the underlying neurophysiological mechanisms. Jefferies *et al*. [43] investigate the neural bases of retrieving knowledge about objects and events, a process that precedes and feeds into composition proper. Meaning composition is likely to occur differently, in mechanistic terms, depending on whether the material to be composed is retrieved in a context- or task-sensitive versus -insensitive way. In particular, context or task sensitivity requires *control processes*, and semantic combinatorics may indeed depend on the extent to which such processes are engaged in a particular context or task, and on how they interact with stored knowledge. Jefferies *et al*. [43] propose that (relatively) uncontrolled retrieval of coherent semantic knowledge engages anterior regions of the temporal lobe, while controlled processes would additionally recruit posterior temporal and inferior frontal cortices. Pylkkänen [44] discusses MEG experiments on the role of the anterior temporal lobe (ATL) in semantic processing. She argues that left ATL responses are sensitive to subtle conceptual semantic relations between words, in ways that are not predicted by the standard account of composition in formal semantics. Pylkkänen's [44] perspective raises the issue whether neural correlates of syntactic and logico-semantic composition—beyond conceptual combination—may/will be identified, and if not, what that means for the role of syntax and logic in the architecture of language. Hagoort [45] presents a model of dynamic interaction between temporo-parietal and inferior frontal cortex. He emphasizes the fact that language interpretation typically happens in rich conversational settings, which provide multi-modal information cues for the construction of models that go beyond a syntax-driven combination of lexical meanings. A neurobiology of linguistic meaning should address, also through mechanistic models, this *non-compositional* process of meaning construction. Calmus *et al*. [46] develop a neurocognitive account of combinatorial binding that aims to explain how (hierarchical) dependency relations are recovered from serial order in (meaningful) sequences. Their account focuses on interactions between inferior frontal and temporal cortical structures, striving to capture experimental data from structured sequence learning tasks.

The second theme is *computational models of composition*. Five contributions are included in this set, representing a broad spectrum of approaches to the algorithmic and neural implementation of semantics. Vankov & Bowers [47] show that neural networks with a specific architecture that allows binding of fillers to roles, given a particular training routine that presses the system to learn to encode semantic relations flexibly, can achieve combinatorial generalization. Baroni [48] considers the performance of current deep neural networks on tasks involving generalization. He shows that, while these systems may be capable of structure-dependent generalizations, also from linguistic data, they do not display significant systematic compositionality. Baroni [48] concludes that further work is needed to gain novel insights on the mechanisms that deep networks could be employing in these tasks. Also, he recommends experimenting with alternative cognitive architectures that can directly support compositionality. Martin & Doumas [49] present one such possible model, which takes seriously the requirement of independence of predicates and arguments (or variables and values) and the need for making explicit the neurophysiological mechanisms underlying composition. They show that models that use tensor products for binding violate

the independence condition, while human behaviour does not, and propose an alternative binding mechanism, based on oscillatory activity driven by inhibitory signals in a settling neural network, that fully preserves independence of predicates and arguments. Gwilliams [50] presents a formal and cognitive model of morphological composition, or how lexical and functional morphemes are composed into known and new words. The model includes a rule-based *composition* step, where morpheme meanings are combined into word meanings, followed by an *update* step that uses composition outputs at the sentence level to adjust properties of individual morphemes. Rabovsky & McClelland [51] focus on (quasi-)compositional representations of meaning in the Sentence Gestalt model, reassessing the model's empirical coverage vis-à-vis language-related ERP effects, such as the N400 and P600. They argue that an artificial neural network model that does not encode sentence meaning fully compositionally could nonetheless capture interesting aspects of lexical and sentential semantics—an argument that resonates with Baroni's [48] conclusion. The five articles in this set touch upon limitations of connectionist and deep neural networks in capturing compositional meaning, but they also highlight some of the conditions under which such models may successfully reproduce aspects of semantic processing. Most importantly, these papers remind us of the importance of models with *interpretable functional states*. Contemporary artificial neural networks, and more specifically deep learning networks, only allow the modeller to 'read off' states from output layers: however there is no principled way of interpreting the network's *internal states*.

The third theme is *detecting and characterizing neural signatures of semantic composition in known or new signal domains* in experimental data. Nieuwland *et al*. [52] present a large-scale ($N = 334$) multi-lab event-related potential (ERP) experiment addressing the long-running debate on the N400 as a signature of *lexical semantic activation*, driven by word predictability, versus the N400 as an index of *semantic integration*, driven by sentence plausibility. Their ERP results show that the amplitude of the N400 component can be modulated by both factors, but that the effects of predictability are observed *before* those of plausibility. Meaning composition can be facilitated by lexical predictability and sentence plausibility, and those facilitatory effects are reflected in attenuations of the N400 amplitude. These results support multiple-generator, multiple-process models of N400 activity, emphasizing the neurophysiological complexity of the N400 response as well as the computational continuity between semantic activation and unification of lexical meaning in context [53,54]. Brennan & Martin [55] use existing EEG data on naturalistic story listening to study phase alignment in correspondence to the onsets of words that close syntactic phrases. They show increased phase synchronization across several frequency bands depending on the number of phrases completed by a word. This finding provides preliminary empirical grounding for some claims made in other papers in this theme issue, such as Martin & Doumas [49]. Fyshe [56] presents an elaboration and application of the Temporal Generalization Method (TGM; [57]) to the detection of traces of compositional processing in brain data. The TGM allows one to assess whether neural activity patterns, at any given time, may contain information on previous activity states: this is especially useful for studying composition, which requires the (re)activation and maintenance of previously processed meanings given the meaning of the current word. The contributions in this set also showcase the richness of EEG data, and assert the need to harness multiple signal domains in EEG to study semantic processes, including lexical semantic activation and composition.

The fourth theme is the *formal and cognitive foundations of compositionality*. Phillips [58] analyses compositionality in formal and cognitive systems within the mathematical framework of category theory. He argues that sheafs (universal morphisms) capture, at the most abstract level of description, the process of constructing a globally coherent representation from locally structured data. His approach is also interesting from a philosophical stance, as it reintegrates symbolic (algebraic) and sub-symbolic (geometric) models of composition at a higher level of mathematical abstraction. Moro [59] analyses from the viewpoint of generative syntax the basic structures that carry compositional meaning in natural language, focusing on clause structure. He proposes a configurational derivation of predication as the merging of two symmetrical phrases. He also discusses the implications of this view for the analysis of symmetry-breaking phenomena in syntax (e.g. movement) as well as for neurolinguistics. Finally, Hendriks [60] proposes that compositionality arises as a constraint on meaning in tasks that require coordination between speakers and hearers. She applies insights from optimality theory, and empirical data on asymmetries between comprehension and production in child language use, to derive the principle of compositionality from perspective taking: sentences may be guaranteed to have compositional meaning only when speakers take listeners into account. Together, these papers present novel perspectives on compositionality, going beyond the standard methodological view often adopted in formal semantics [8,16].

This theme issue aims at achieving convergence, and ultimately some degree of integration, between different formal approaches and experimental results on semantic composition, as illustrated by the four themes presented here.

4. Outlook and conclusion

Understanding meaning composition in brains and machines requires a shift in current theory and modelling practice from developing formalisms that can capture properties of composition in logical terms (e.g. the lambda calculus) to constructing models that explain both the human capacity for composition and behavioural and neural data from cognitive processing experiments. Much progress has been made in recent years, as is testified by the contributions in this theme issue, but much remains to be done toward mechanistic models of composition that meet the desiderata discussed here.

For cognitive neuroscience and the neurobiology of language, the challenge is to use correlational spatio-temporal data to identify the cortical networks and neurophysiological events that are *causally responsible* for composition. This requires experimental designs constrained by semantic theory, and ideally by computational models of semantics, in order to determine precisely what the observed brain signals may be neural correlates of. This endeavour is crucial for the neurobiology of language as a whole. Indeed, discounting sensory and perceptual contributions to language processing (decoding auditory or visual or tactile inputs), much of what happens in the brain in response to language is *semantic and pragmatic processing*, with syntax playing a lesser role than or perhaps a different role from that envisaged by the view of syntax–semantics relations in formal semantics and generative syntax [5,61].

For linguistics and computer science, the challenge is to develop models that can address compositional *and* non-compositional aspects of meaning, using reasonable definitions of compositionality that make the principle non-trivial or non-vacuous formally and empirically; this specific task is likely to require incisive contributions from philosophers of language and mind. For example, current connectionist or deep learning models capture experience-based and often non-compositional aspects of meaning, and can behave compositionally only with pre-wired architectures or under specific training regimes or given carefully crafted training data. The issue, then, is whether compositionality is a sensible benchmark for connectionist or deep neural network models. This stands in contrast with the capabilities of neural–symbolic models, which can support compositionality via independent role and filler representations, but may be less sensitive to subtle variations of meaning across contexts or over time, unless they are also based in distributed representations. This apparent divide may be resolved by neurocognitive architectures where compositional and non-compositional processing are allocated to distinct semi-autonomous 'modules'—respectively, a symbolic, syntax-driven component that composes lexical meanings productively, complying with the independence of roles and fillers, and a non-symbolic, context-driven component that relates meanings predictively, tracking statistical or any other relevant regularities in the data. However, any hybrid architecture must possess the ability to discover, learn, and deploy new symbols from unstructured data, else it will face the problem that Bayesian program induction and most other hard-coded symbol systems presently face, i.e. the catch-22 requirement of knowing *a priori* the solution (e.g. the graph of the representation system, not merely the training data; see [6,62–64]). Hybrid architectures of this kind may prove the most successful in the long run if this problem is solved; at present, they remain unique in their ability to provide an overarching vista on theoretical, computational and experimental research on syntactic and semantic composition.

Data accessibility. This article has no additional data.

Authors' contributions. A.E.M. and G.B. wrote the paper.

Competing interests. We declare we have no competing interests.

Funding. We gratefully acknowledge funding from the Research Council of Norway to G.B. (FRIHUMSAM 251219) and from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) (016.Vidi.188.029) and the Max Planck Society (Independent Research Group 'Language and Computation in Neural Systems') to A.E.M. We are grateful to the Faculty of Humanities at NTNU, the Norwegian University of Science and Technology, for co-funding the international symposium 'Towards mechanistic models of meaning composition', 11–12 October 2018, which led to this theme issue.

References

1. Partee B. 1984 Compositionality. In *Varieties of Formal Semantics: Proc. 4th Amsterdam Colloquium* (eds F Landman, F Veltman), *September 1982*, pp. 281–311. Dordrecht, The Netherlands: Foris Publications.
2. Janssen TM. 1997 Compositionality. In *Handbook of logic and language* (eds JFAK van Benthem, A ter Meulen), pp. 417–473. Amsterdam, The Netherlands: North-Holland.
3. Fodor JA, Pylyshyn ZW. 1988 Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71. (doi:10.1016/0010-0277(88)90031-5)
4. Martin AE. 2016 Language processing as cue integration: grounding the psychology of language in perception and neurophysiology. *Front. Psychol.* **7**, 120. (doi:10.3389/fpsyg.2016.00120)
5. Baggio G. 2018 *Meaning in the brain*. New York, NY: MIT Press.
6. Martin AE, Doumas LA. 2019 Predicate learning in neural systems: using oscillations to discover latent structure. *Curr. Opin. Behav. Sci.* **29**, 77–83. (doi:10.1016/j.cobeha.2019.04.008)
7. Westerståhl D. 1998 On mathematical proofs of the vacuity of compositionality. *Ling. Philos.* **21**, 635–643. (doi:10.1023/A:1005401829598)
8. Groenendijk JAG, Stokhof MJB. 2004 Why compositionality? In *The Partee effect* (eds GN Carlson, J Pelletier), pp. 83–106. Stanford, CA: CSLI Press.
9. Pagin P, Westerståhl D. 2010 Compositionality I: definitions and variants. *Philos. Compass* **5**, 250–264. (doi:10.1111/j.1747-9991.2009.00228.x)
10. Pagin P, Westerståhl D. 2010 Compositionality II: arguments and problems. *Philos. Compass* **5**, 265–282. (doi:10.1111/j.1747-9991.2009.00229.x)
11. Dowty D. 2007 Compositionality as an empirical problem. In *Direct compositionality* (eds C Barker, P Jacobson, PI Jacobson), pp. 23–101. Oxford, UK: Oxford University Press.
12. van Rooij I. 2008 The tractable cognition thesis. *Cogn. Sci.* **32**, 939–984. (doi:10.1080/03640210801897856)
13. Montague R. 1970 English as a formal language. In *Linguaggi nella società e nella tecnica* [*Languages in society and technology*] (ed. B Visentini), pp. 189–223. Milan, Italy: Edizioni di Comunità.
14. Partee BH, ter Meulen A, Wall RE. 2012 Mathematical methods in linguistics, *Studies in Linguistics and Philosophy*, vol. 30. Dordrecht, The Netherlands: Kluwer Academic Publishers.
15. Heim I, Kratzer A. 1998 *Semantics in generative grammar*. Oxford, UK: Blackwell.
16. Baggio G, van Lambalgen M, Hagoort P. 2012 The processing consequences of compositionality. In *The Oxford handbook of compositionality* (eds W Hinzen, E Machery, M Werning), pp. 657–674. Oxford, UK: Oxford University Press.
17. Yalcin S. 2014 Semantics and metasemantics in the context of generative grammar. In *Metasemantics: new essays on the foundations of meaning* (eds A Burgess, B Sherman), pp. 334–360. Oxford, UK: Oxford University Press.
18. Katz JJ, Fodor JA. 1963 The structure of a semantic theory. *Language* **39**, 170–210. (doi:10.2307/411200)
19. Geeraerts D. 2010 *Theories of lexical semantics*. Oxford, UK: Oxford University Press.
20. Jackendoff R. 1983 *Semantics and cognition*. New York, NY: MIT Press.
21. Pustejovsky J. 1995 *The generative lexicon*. New York, NY: MIT Press.
22. Harris ZS. 1954 Distributional structure. *Word* **10**, 146–162. (doi:10.1080/00437956.1954.11659520)
23. Mitchell J, Lapata M. 2010 Composition in distributional models of semantics. *Cogn. Sci.* **34**, 1388–1429. (doi:10.1111/j.1551-6709.2010.01106.x)
24. Baroni M. 2013 Composition in distributional semantics. *Lang. Ling. Compass* **7**, 511–522. (doi:10.1111/lnc3.12050)
25. Jackendoff R. 2002 *Foundations of language: brain, meaning, grammar, evolution*. Oxford, UK: Oxford University Press.
26. Marr D. 1982 *Vision: a computational investigation into the human representation and processing of visual information*. New York, NY: Henry Holt & Co.
27. Reverberi C, Görgen K, Haynes JD. 2012 Compositionality of rule representations in human prefrontal cortex. *Cereb. Cortex* **22**, 1237–1246. (doi:10.1093/cercor/bhr200)
28. RE Hume. 1921 *Chandogya Upanishad. The thirteen principal Upanishads: translated from the Sanskrit with an outline of the philosophy of the Upanishads and an annotated bibliography*, §18. London, UK: Oxford University Press.
29. Hock HH. 2005 Philology and the historical interpretation of the Vedic texts. In *The Indo-Aryan controversy: evidence and inference in Indian history*

(eds E Bryant, L Patton), pp. 282–308. Oxford, UK: Routledge Taylor & Francis Group.
30. Hempel CG, Oppenheim P. 1948 Studies in the logic of explanation. *Philos. Sci.* **15**, 135–175. (doi:10.1086/286983)
31. Newell A. 1973 You can't play 20 questions with nature and win: projective comments on the papers of this symposium. In *Visual Information Processing: Proc. 8th Annu. Carnegie Symp. Cognition, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 19 May, 1972* (ed. WG Chase), pp. 1–26. New York, NY: Academic Press.
32. Kaplan DM. 2011 Explanation and description in computational neuroscience. *Synthese* **183**, 339. (doi:10.1007/s11229-011-9970-0)
33. Bechtel W. 2012 *Mental mechanisms: philosophical perspectives on cognitive neuroscience*. New York, NY: Routledge.
34. Kaplan DM, Craver CF. 2011 The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective. *Philos. Sci.* **78**, 601–627. (doi:10.1086/661755)
35. Bechtel W, Abrahamsen A. 2005 Explanation: a mechanist alternative. *Stud. Hist. Philos. Sci. C* **36**, 421–441. (doi:10.1016/j.shpsc.2005.03.010)
36. Glennan S. 2017 *The new mechanical philosophy*. Oxford, UK: Oxford University Press.
37. Levy A, Bechtel W. 2013 Abstraction and the organization of mechanisms. *Philos. Sci.* **80**, 241–261. (doi:10.1086/670300)
38. Salmon WC. 1984 *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
39. Craver CF. 2002 Structures of scientific theories. In *The Blackwell guide to the philosophy of science* (ed. P Machamer, M Silberstein), pp. 55–79. Oxford, UK: Blackwell.
40. Piccinini G. 2007 Computing mechanisms. *Philos. Sci.* **74**, 501–526. (doi:10.1086/522851)
41. Miłkowski M. 2013 *Explaining the computational mind*. Cambridge, MA: MIT Press.
42. Van Rooij I. 2008 The tractable cognition thesis. *Cogn. Sci.* **32**, 939–984. (doi:10.1080/03640210801897856)
43. Jefferies E, Thompson H, Cornelissen P, Smallwood J. 2019 The neurocognitive basis of knowledge about object identity and events: dissociations reflect opposing effects of semantic coherence and control. *Phil. Trans. R. Soc. B* **375**, 20190300. (doi:10.1098/rstb.2019.0300)
44. Pylkkänen L. 2019 Neural basis of basic composition: what we have learned from the red–boat studies and their extensions. *Phil. Trans. R. Soc. B* **375**, 20190299. (doi:10.1098/rstb.2019.0299)
45. Hagoort P. 2019 The meaning-making mechanism(s) behind the eyes and between the ears. *Phil. Trans. R. Soc. B* **375**, 20190301. (doi:10.1098/rstb.2019.0301)
46. Calmus R, Wilson B, Kikuchi Y, Petkov CI. 2019 Structured sequence processing and combinatorial binding: neurobiologically and computationally informed hypotheses. *Phil. Trans. R. Soc. B* **375**, 20190304. (doi:10.1098/rstb.2019.0304)
47. Vankov II, Bowers JS. 2019 Training neural networks to encode symbols enables combinatorial generalization. *Phil. Trans. R. Soc. B* **375**, 20190309. (doi:10.1098/rstb.2019.0309)
48. 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)
49. Martin AE, Doumas LAA. 2019 Tensors and compositionality in neural systems. *Phil. Trans. R. Soc. B* **375**, 20190306. (doi:10.1098/rstb.2019.0306)
50. Gwilliams L. 2019 How the brain composes morphemes into meaning. *Phil. Trans. R. Soc. B* **375**, 20190311. (doi:10.1098/rstb.2019.0311)
51. 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. (doi:10.1098/rstb.2019.0313)
52. Nieuwland MS *et al.* 2019 Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Phil. Trans. R. Soc. B* **375**, 20180522. (doi:10.1098/rstb.2018.0522)
53. Pylkkänen L, Marantz A. 2003 Tracking the time course of word recognition with MEG. *Trends Cogn. Sci.* **7**, 187–189. (doi:10.1016/S1364-6613(03)00092-5)
54. Baggio G, Hagoort P. 2011 The balance between memory and unification in semantics: a dynamic account of the N400. *Lang. Cogn. Processes* **26**, 1338–1367. (doi:10.1080/01690965.2010.542671)
55. Brennan JR, Martin AE. 2019 Phase synchronization varies systematically with linguistic structure composition. *Phil. Trans. R. Soc. B* **375**, 20190305. (doi:10.1098/rstb.2019.0305)
56. Fyshe A. 2019 Studying language in context using the temporal generalization method. *Phil. Trans. R. Soc. B* **375**, 20180531. (doi:10.1098/rstb.2018.0531)
57. King JR, Dehaene S. 2014 Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210. (doi:10.1016/j.tics.2014.01.002)
58. Phillips S. 2019 Sheaving—a universal construction for semantic compositionality. *Phil. Trans. R. Soc. B* **375**, 20190303. (doi:10.1098/rstb.2019.0303)
59. Moro A. 2019 The geometry of predication: a configurational derivation of the defining property of clause structure. *Phil. Trans. R. Soc. B* **375**, 20190310. (doi:10.1098/rstb.2019.0310)
60. Hendriks P. 2019 The acquisition of compositional meaning. *Phil. Trans. R. Soc. B* **375**, 20190312. (doi:10.1098/rstb.2019.0312)
61. Pylkkänen L. 2019 The neural basis of combinatory syntax and semantics. *Science* **366**, 62–66. (doi:10.1126/science.aax0050)
62. Doumas LAA, Hummel JE. 2005 Approaches to modeling human mental representations: what works, what doesn't and why. In *The Cambridge handbook of thinking and reasoning* (eds KJ Holyoak, RG Morrison), pp. 73–94. Cambridge, UK: Cambridge University Press.
63. Doumas LAA, Hummel JE, Sandhofer CM. 2008 A theory of the discovery and predication of relational concepts. *Psychol. Rev.* **115**, 1–43. (doi:10.1037/0033-295X.115.1.1)
64. Holyoak KJ, Hummel JE. 2000 The proper treatment of symbols in a connectionist architecture. *Cogn. Dyn. Concept. Change Hum. Mach.* **229**, 263.