



Neurobiological causal models of language processing


Hartmut Fitz^{1,2}, Peter Hagoort^{1,2}, and Karl Magnus Petersson^{2,3}

¹Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands

²Neurobiology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

³Faculty of Medicine and Biomedical Sciences, University of Algarve, Faro, Portugal

Author Note

Hartmut Fitz  <https://orcid.org/0009-0000-8821-2312>

Correspondence concerning this article should be addressed to Hartmut Fitz, Donders Institute for Brain, Cognition and Behaviour, Thomas van Aquinostraat 4, 6525 GD Nijmegen, the Netherlands. Email: hartmut.fitz@donders.ru.nl

Abstract

The language faculty is physically realized in the neurobiological infrastructure of the human brain. Despite significant efforts, an integrated understanding of this system remains a formidable challenge. What is missing from most theoretical accounts is a specification of the *neural mechanisms* that implement language function. Computational models that have been put forward generally lack an explicit neurobiological foundation. We propose a neurobiologically informed causal modeling approach which offers a framework for how to bridge this gap. A *neurobiological causal model* is a mechanistic description of language processing that is grounded in, and constrained by, the characteristics of the neurobiological substrate. It intends to model the *generators* of language behavior at the level of implementational causality. We describe key features and neurobiological component parts from which causal models can be built and provide guidelines on how to implement them in model simulations. Then we outline how this approach can shed new light on the core computational machinery for language, the long-term storage of words in the mental lexicon and combinatorial processing in sentence comprehension. In contrast to cognitive theories of behavior, causal models are formulated in the ‘machine language’ of neurobiology which is universal to human cognition. We argue that neurobiological causal modeling should be pursued in addition to existing approaches. Eventually, this approach will allow us to develop an explicit computational neurobiology of language.

Keywords: neurobiology of language, computational modeling, implementational causality, adaptive dynamical systems, processing memory

Neurobiological causal models of language processing

“You can’t go to a physics conference and say: I’ve got a great theory. It accounts for everything and is so simple it can be captured in two words: *Anything goes*”—Noam Chomsky¹

The core computational machinery for language

Sentence comprehension requires at least two functional components, a long-term storage of words and their feature structure (mental lexicon) and a combinatorial device (unification) that integrates sequential information into structured representations over time (Hagoort, 2005, 2019; Jackendoff, 2002). These components interact during real-time, incremental processing and mutually control each other. This process involves linguistic representations at different grain sizes, from phonemes to words, phrases and sentences (Dehaene et al., 2015), and memory on multiple **timescales** (see Glossary), ranging from milliseconds to minutes and a lifetime (Hasson et al., 2015). An adaptive processing dynamics shaped by ontogenetic development (genes and experience) operates on these linguistic primitives and ties them together in processing memory (Petersson & Hagoort, 2012). The computational machinery that supports these operations is implemented in neurobiological infrastructure at different spatial scales, from single neurons and synapses to cortical layers, micro-columns, brain regions and large-scale networks. A theory of language processing that aims to be *complete* needs to explain how this machinery is realized within the neurobiology of the language system² across spatial and temporal scales. This explanatory goal is shared by most researchers in the field, but an integrated account has not been accomplished thus far. Some have argued that we lack even the most basic understanding of how linguistic units are represented and stored in long-term memory (Poeppel & Idsardi, 2022). In a similar vein, the neurobiological basis of processing memory for unification is currently unknown (Fields, 2022; Fitz et al., 2020). In this perspective article, we describe a computational modeling

¹ <https://garymarcus.substack.com/p/noam-chomsky-and-gpt-3>

² We use the term ‘language system’ as short-hand for ‘language-relevant brain regions’ without implying that these regions are functionally exclusive to language.

approach that maps out a way forward for the language sciences in order to achieve this explanatory goal. This approach aims towards a fundamental understanding of core language function from first principles of neurobiology.

Multiple explanatory strategies

Language as a neurobiological system needs to be distinguished from its behavioral output which includes speech, sign or text in production and sentence interpretations in comprehension. Although the language system is used for communication and thinking, these phenomena should not be mistaken for the system itself (Jackendoff, 2002). A key question is how to link behavioral output to the computational machinery of the neurobiological system that generates the output. This is one of the fundamental challenges in explaining natural language in mechanistic terms.

The experimental approach sets out at the functional level of description³ and attempts to infer processing theories from measured input-output relations (Figure 1). These are often informal verbal theories that do not reach algorithmic specificity. Moreover, current experimental methods are relatively coarse and do not allow the reconstruction of simple computational devices whose functionality is known (Jonas & Kording, 2017). This complicates the reverse engineering of cognitive systems from experimental data which therefore has to be complemented with other methods. One such approach has tried to map these relations algorithmically through cognitive modeling. Different frameworks have been proposed (e.g., connectionist, symbolic, hybrid, Bayesian, etc.) that each captures some aspect of language behavior, but so far this approach has not resulted in a unified picture of linguistic computation. Since any finite collection of data can be re-coded by many different formalisms, success in approximating behavior algorithmically does not automatically guarantee neurobiological realism. The chances that a stipulated algorithm provides a correct description of the actual computational machinery is small, no matter how well the formalism fits with behavioral data. Independent evidence is needed to establish realism which, by necessity, must stem from the neurobiological characteristics of the very system that is

³ The top level refers to *what* is being computed, i.e., which recursive function ϕ . We therefore label it ‘functional level’. Marr’s term ‘computational level’ is unfortunate because it creates confusion with ‘algorithm’.

being modeled. In the absence of such neurobiological constraints, cognitive models remain high-level abstractions whose relationship to the implementational substrate is unclear.

For these reasons, we argue that a third explanatory strategy should be pursued urgently, and concurrent with the more traditional approaches shown in Figure 1. This strategy puts a premium on neurobiology as a primary source of evidence and attempts to model the language system at the implementational level of description. We refer to this approach as ‘neurobiological causal modeling’.⁴ A causal model is built directly on established neurobiological principles without making *ad hoc* assumptions about algorithmic procedures and component parts (Figure 2). The goal of this approach is to synthesize an explanatory language model that can uniformly explain linguistic behavior across different experiments. Unlike most existing approaches, causal modeling draws on a wealth of additional insights from neuroanatomy (Petrides, 2014; Tremblay & Dick, 2016), neurophysiology (Kandel et al., 2012; Luo, 2015; Sterling & Laughlin, 2015) and biophysics (Koch, 1999) that inform model construction. The implementational building blocks derived from these knowledge sources can provide the necessary constraints for a computational neurobiology of language that ultimately integrates across all levels of description.

The case for neurobiological constraints on models of cognition was made in the seminal work of Churchland and Sejnowski (1992) and has been reiterated by others since then (e.g., Astle et al., 2023; O’Reilly, 2006; Pulvermüller et al., 2021). One way to approach this issue is to constrain existing neurocognitive architectures in order to increase their biological plausibility (Pulvermüller et al., 2021). Another approach, which we advocate here, is to systematically assemble computational language models from known neurobiological primitives at the implementational level (Figure 2). Although superficially similar, the former approach is reductive in nature while the latter is synthetic. To prevent early misunderstanding, neurobiological causal language modeling does not strive to dispense with function or algorithm

⁴ In our terminology, a causal model is a set of functional equations that describes the dynamics of a system at the level of neurobiological causality. This concept differs from the structural causal models of Pearl (2000), dynamic causal modeling (Friston et al., 2003), or models of causality itself (Granger, 1969).

which are an integral part of a complete explanation. On the contrary, causal modeling aims to firmly ground linguistic behavior and cognitive theory in the causal characteristics of the actual language system and its concrete neurobiological instantiation.

First principles of neurobiology

The language system of the human brain is a particular instance of a sparsely connected recurrent network of biological neurons and chemical synapses. This theoretical framework is sufficiently expressive to capture *all* anatomical connectivity, including connectivity between brain regions, the **laminar structure** within cortical columns, **synaptic motifs** within and between layers, and randomness at the microscopic scale. In the context of recurrent networks, there is no fundamental difference between connectivity patterns at different spatial scales. Since a static structured connectome by itself is non-explanatory (Bargmann, 2012), it is critical to also realistically model neural interactions and the information flow across this graph.

Fast signaling in the nervous system is based on **action potentials** which are all-or-none neuronal responses to analog input. Spikes are the basic units of cortical information processing and it has been argued that their temporal relations play an important role in the encoding, representation and transmission of processing outcomes (Brette, 2015; Gerstner et al., 1997). Neurobiological language models are needed that can express the temporal dimension of spike-based processing and resolve the mismatch between the timescales of action potentials and cognitive behavior (Chaudhuri & Fiete, 2016). Biological neurons exhibit a wide range of electrophysiological behavior, from tonic spiking to bursting and adaptation, and this diversity of observed firing patterns is likely to have functional significance (Gerstner et al., 2014; Koch, 1999). Neuronal spike responses result from the integration of synaptic inputs on the spatial structure of the **dendritic** tree which amounts to more than linear summation. The spatio-temporal nature of dendritic integration gives rise to complex, non-linear processing effects that are not captured by simpler **point neurons** (Gidon et al., 2020; London & Häusser, 2005; Payeur et al., 2019). Thus, the input–output behavior of neurons as the fundamental computational unit is substantially richer than has been assumed (Larkum, 2022). Dendritic

morphology is one of the candidate features that may account for species-specific cognitive functions (Fişek & Häusser, 2020), including language, and multi-compartment neuron models can be viewed as interconnected computational elements that are all potential targets for learning and adaptation (see the Tripod neuron for an explicit modeling account, Quaresima et al., 2022).

Neurons connect via excitatory or inhibitory synapses but not both at the same time and synapses do not change sign during learning and development (Strata & Harvey, 1999), as is the case in virtually all connectionist and deep learning models of language processing. Major synapse types include fast and slow excitatory and inhibitory ones that generate post-synaptic currents with different polarity, amplitudes, and rise and decay timescales (Destexhe et al., 1998). Synaptic learning and memory are subserved by a variety of **unsupervised learning** principles (Magee & Grienberger, 2020) that include activity-dependent, short-term synaptic changes (Markram et al., 1998), mechanisms for long-term potentiation and depression based on the timing of pre- and post-synaptic spikes (Markram et al., 1997), as well as synaptic consolidation on much longer timescales (Clopath, 2012). In addition, reward-modulated learning (Frémaux & Gerstner, 2016) and more powerful error-driven learning mechanisms also play a role (Payeur et al., 2021; Whittington & Bogacz, 2019).

To temper runaway processes due to Hebbian plasticity, **homeostatic** mechanisms need to ensure that single-neuron and circuit firing rates remain within physiological ranges (Tetzlaff et al., 2012; Turrigiano & Nelson, 2004). These mechanisms act, e.g., by scaling synaptic **conductances** or by downregulating neuronal excitability. Furthermore, language-relevant networks need to function in the presence of endogenous background activity and stochastic variability at the cellular and synaptic level (Faisal et al., 2008; Nolte et al., 2019). These noise sources reduce the computational capacity of the system to that of Turing machines with finite tapes, i.e., finite-state machines, by limiting processing precision and effective memory capacity (Maass & Orponen, 1998; Petersson, 2005). In addition, we note that in parallel with the fast processing systems outlined above, there are neuromodulatory systems (e.g., **monoamines**, **neuropeptides**, etc.) that are different in nature from the fast conductance-based signaling

systems. They typically originate in the midbrain/brainstem, with widespread cortical–subcortical projections, operate on longer timescales, and directly regulate the intracellular biochemistry via **G protein** coupled receptors. These systems modulate fast neural processing and it has been suggested that they support unconventional computation and neuronal memory (Bechtel, 2022; Bray, 2009; Koch, 1999).

This inventory of neurobiological principles constitutes the foundation of causal modeling and imposes strong constraints on the computational realization of language (Figure 2). Importantly, these constraints are both constructive and limitative. On the one hand, they specify the basic building blocks of neurobiological language models and thus provide an evidence-based implementational scaffold for causal modeling. Mathematical models of these component parts have been carefully developed by experimental and theoretical neuroscientists to closely capture the net effects of physiological processes quantitatively (Box 1). The objective of causal modeling is to explain language processing in terms of these neurobiological principles that characterize the mechanics of the real system. On the other hand, these constraints curb arbitrary choices made in cognitive language modeling and deep learning models at the level of component parts and algorithms. In order to establish *valid* abstractions, it is necessary to scientifically demonstrate that these abstractions can be reduced to the level of neurobiological implementation. Pending such reductions, algorithmic explanations that are obtained by abstracting away from elementary features of the nervous system run a high risk of being spurious.

Dynamical systems view on language

The neurobiology of language fits naturally within a description of language processing in terms of a specific continuous-time adaptive dynamical system built from neurobiological components. Here we provide a terse mathematical formalization of such a system \mathcal{S} in terms of interacting functional components that are coupled via a neurobiologically specified processing dynamics \mathcal{P} and adaptive learning mechanisms \mathcal{L} (Figure 3). Note that \mathcal{P} and \mathcal{L} are multivariate and each component is associated with a physical measurement unit.

\mathcal{P} maps an internal state $s \in \Omega$ and an input $i \in \Sigma$ onto a new internal state

$\hat{s} = s(t + dt) = s(t) + ds(t)$. States s are real-valued tuples of **dynamical variables** in neurobiology, e.g., membrane potentials and synaptic conductances, that describe the language system across all spatial scales. Input i is provided to \mathcal{P} by the environment that the system is embedded in through an interface (e.g., a speech sound transduced by the cochlea) and the optional output λ of \mathcal{P} is translated into an internal action or an external motor response (e.g., articulation). State transition is characterized by coupled stochastic differential equations $ds(t) = \mathcal{P}(s, i, m)dt + d\xi(t)$ that are parameterized in m (see below) and coupled to noise processes $\xi(t)$. Thus, information processing is represented as an input-driven, or *forced*, trajectory through the system's state space Ω and, importantly, is constrained by the dynamics \mathcal{P} . In classical terminology this is the infinitesimal version of the process logic of a Turing-machine, i.e., its machine or transition table. For instance, in language comprehension, a subsystem of \mathcal{P} can be understood as the parser associated with \mathcal{S} . Since $ds(t)$, and therefore the next state \hat{s} , is recursively determined by the continuous action of \mathcal{P} , language processing in this framework is naturally incremental, recursive and state-dependent, as in classical theories of computation (Buonomano & Maass, 2009; Petersson & Hagoort, 2012).

\mathcal{P} is intertwined with a dynamics \mathcal{L} for development, learning and adaptation that governs the evolution of \mathcal{S} as a function of linguistic experience and maturation. This is formalized as $dm(t) = \mathcal{L}(m, s, t)dt + d\eta(t)$ where the learning parameters m belong to the model space $\mathbf{M} = \{m \mid m \text{ can be realized by } \mathcal{S}\}$ and $\eta(t)$ is another noise process. The elements of \mathbf{M} are high-dimensional tuples of synaptic, neuronal and other adaptive parameters in the language network, and the dynamics \mathcal{L} is a set of neurobiological learning principles. In contrast to \mathcal{P} , \mathcal{L} is explicitly dependent on time \mathbf{T} which captures the notion of innately guided maturation processes. At any point in time, \mathcal{S} is in a particular developmental state $m(t)$ and \mathcal{L} carves out a trajectory in \mathbf{M} as the system matures. However, since \mathcal{L} is coupled back to \mathcal{P} via m , the processing characteristics of \mathcal{S} themselves change over time, and the fixed points of \mathcal{L} mark the developmental end-state of adult competence. Prior knowledge of language (Chomsky, 1986) is incorporated into \mathcal{S} as a structured initial state $m(t_0)$, or as additional constraints on \mathcal{P} , \mathcal{L} , or \mathbf{M} ,

the so-called language acquisition device (cf. Petersson & Hagoort, 2012). The initial state is the outcome of gene-regulatory development of the language-ready brain, optimized by biological evolution, and subsequently fine-tuned through linguistic experience during acquisition (Zador, 2019). Due to the fact that the complete dynamics of \mathcal{S} is also shaped by linguistic interaction with a cultural environment, the neurobiological language system is a biocultural hybrid (Evans & Levinson, 2009).

Consequently, the general form of the language system \mathcal{S} is an adaptive system of interacting dynamical variables in neurophysiology whose state transitions are determined by the coupled dynamics for processing \mathcal{P} and learning/development \mathcal{L} . At any developmental stage, the algorithmic nature of \mathcal{P} and \mathcal{L} is determined by neurobiology and one objective of causal modeling is to characterize these dynamics and interpret them in language processing terms. Without a cognitive interpretation, \mathcal{S} remains an unanalyzed system that moves in time. Another important goal of causal modeling is to identify the language-relevant representational states of \mathcal{S} which are expected to be evoked spatio-temporal transients in ongoing processing (Petersson, 2008; Rabinovich et al., 2008).

The dynamical systems perspective characterizes language processing in full generality and with formal precision. This allows us to clearly identify the different *explananda*—processing, learning, maturation and the initial state—and how they interact. Component parts of causal models are expressed as continuous-time differential equations coupled into a functional architecture defined by the connectome. Every instantiation of a causal language model built from such component parts *ipso facto* is a specific claim about, and a concrete algorithmic proposal of how, the processing and learning dynamics \mathcal{P} and \mathcal{L} could be implemented at the level of neurobiology. Hence, there is a natural relationship between the *neurobiological* dynamical system and causal language modeling whereas this link is either missing or contrived for models that are not formulated in causal terms.

Hierarchy and binding in neural processing

Language is characterized in terms of hierarchical structures that describe the representations that the comprehension system needs to compute when parsing an utterance. Hierarchical dependencies between constituents are ubiquitous at all linguistic levels, from phonemes and syllables to words, phrases, clauses and sentences (Hagoort, 2019; Hasson et al., 2015; Jackendoff, 2002). At the same time, language processing is subserved by recurrent networks of spiking neurons and chemical synapses and it is not obvious how hierarchical linguistic structure can be mapped to neurobiology. Thus it has been an enduring debate how neural systems can accomplish so-called ‘hierarchical processing’ and this issue is closely tied to the binding problem.

The apparent conflict between these notions can be resolved when static structural hierarchy (represented by parse trees) is interpreted dynamically in neural processing terms (Figure 4) where words are retrieved from the mental lexicon by an operator R and unified combinatorially by a universal function U . Hierarchical processing corresponds to nested function calls, including recursion, that are executed by the neural parser at the appropriate point in logical time, augmented with a memory structure, or unification space, to store and retrieve intermediate results when needed. The control input for U parametrically switches unification into different subroutines by function composition. It is supplied by the feature structure of retrieved words (e.g., lexical categories), or computed internally within processing memory from the available information (e.g., phrasal categories). Biological networks for unification thus require distinct input lines for data and control, similar to the pins on a microprocessor. In neurobiology this can be achieved by electrotonically segregated dendritic branches that integrate different input types independently (Larkum, 2022; Spratling, 2002). For example, basal and apical dendrites of cortical pyramidal neurons receive inputs from anatomically distinct source locations that differentially modulate the somatic response (Binzegger et al., 2004; Lafourcade et al., 2022; Sheperd, 2004). This spatial separation of distinct classes of inputs explains how a single neuron (or circuit, for that matter) can play different functional roles in unification, from one time step to

the next.

The translation in Figure 4 shows how to resolve the perceived mismatch between hierarchy and brain networks, going from parse trees to function composition to neural processing. When cast in functional terms, static hierarchical phrase structure trees can be given a dynamic interpretation in terms of recurrent neural processing with the appropriate memory structure. It also shows that hierarchical processing does not require the construction of *explicit* representations of linguistic trees and their binding relations (as some models have suggested, e.g., Martin & Doumas, 2017; Papadimitriou & Friederici, 2022; van der Velde & Kamps, 2006) because these relations are already implicitly present in the intermediate processing outcomes of the state-dependent neural parser. As words are being processed one-by-one, the system incrementally computes an interpretation in neuronal memory registers, i.e., dynamical variables in processing memory, which are a particular sub-state of the complete system state. Parsing “the cat chases a dog” versus “a dog chases the cat” results in distinct trajectories whose end-states represent different meanings. This procedure is analogous to evaluating a hierarchically structured arithmetic expression by a compiled program where the final outcome is a number that corresponds to the correct interpretation, rather than an explicit structural representation of the binary expression tree. Introspection of constituent structure requires linguistic knowledge and should not be considered part of automatic language processing.

Function composition and binding in comprehension rely on data structures that must be supported by neurobiology. The nature of these data structures determines the kind of unification procedures that can run on partial interpretations temporarily held in processing memory. Data structures and how they are represented in memory are a key organizing principle of neurobiological information processing systems. For example, the membrane potential, or other dynamical variables, of a biological neuron assumes real number values. The decimal expansion of these numbers can naturally be interpreted as a stack memory when combined with push and pop operations. These operations can be implemented through multiplication that shifts decimals into (push) or out of (pop) the decimal expansion and there is evidence that single neurons can

accomplish this (Groschner et al., 2022). More broadly, scaling and other operations on dynamical variables can be viewed as generalized push and pop operations.

In classical computability theory (Cutland, 1980), binding is achieved in that variables are physical memory addresses and the stored bit patterns are their current values. Composite data structures are then assembled by computing references to existing memory content, e.g., using pointers. However, since recurrent networks are fully equivalent to the classical notion of computation (Siegelmann, 1999), binding can also be achieved by neural networks. Binding is therefore not a fundamental barrier and it is an empirical question how it is realized within the specific neurobiological memory architecture. For instance, similar to memory in digital computers, any dynamical variable in physiology with a non-zero time constant is stateful and can act as a memory register. Different information sources can be bound in these registers through temporal integration. Whether this form of binding is sufficient to explain language comprehension or whether other complex neurobiological data structures are required is an open issue, and causal models together with experimental work are needed to answer this question.

Outline of a causal language model

Unification instantiates a generic sequence processor that may not be specific to language (Jackendoff & Audring, 2020; Petersson & Hagoort, 2012) and establishes semantic relations between constituents (e.g., *who does what to whom?*) within processing memory (Figure 5). Traditionally, neurobiological short-term memory has been conceptualized as states of persistent neural activity (Fuster & Alexander, 1971; Goldman-Rakic, 1995). Persistent activity can be achieved through cellular bistability (Loewenstein & Sompolinsky, 2003; Zylberberg & Strowbridge, 2017) or attractor dynamics where excitatory feedback enables the replay of information beyond stimulus offset (Barak & Tsodyks, 2014; Durstewitz et al., 2000). Alternatively, short-term memory has been linked to functional connectivity induced by transient changes in synaptic efficacy (Fiebig & Lansner, 2017; Mongillo et al., 2008). These theories can explain maintenance and cued recall but they have not been developed with language in mind. A neurobiological processing memory for language also needs to be able to integrate and transform

internal representations in an online, incremental fashion and actively compute an interpretation from rapid serial input. In addition, this memory system needs to be context-dependent and sensitive to precedence relations between words. Recent modeling work indicates that these requirements are met by neuronal processing memory (Fitz et al., 2020; Rao et al., 2022) which is grounded in the observation that neurons exhibit adaptive changes in excitability as a function of experience (Marder et al., 1996; Turrigiano et al., 1996). This intrinsic plasticity is common in excitatory cortical cells (Gouwens et al., 2019) and adaptive changes can last from milliseconds (Koch, 1999) to seconds (Levy & Bargmann, 2020) and minutes (Titley et al., 2017). Network simulations have shown that neuronal memory can support sentence-level semantic processing and memory span was proportional to the time constant of spike-rate adaptation (Figure 5). The proposed memory mechanism was also suitable to resolve temporary ambiguity and establish binding relations between words and their semantic roles when queried (Fitz et al., 2020; Uhlmann, 2020). It is likely that other factors contribute to neuronal memory as well, including the kinetics of NMDA-receptors (Lisman et al., 1998) and the morphology of dendrites (Papoutsis et al., 2014; Poirazi & Papoutsis, 2020). These two features support the generation of plateau potentials, endowing neurons with dendritic memory that is useful for structured sequence processing on short timescales (Quaresima et al., 2022). These findings from causal modeling illustrate how evidence from neurobiology can generate new hypotheses about the nature of processing memory for language. Non-causal models do not express these cellular and synaptic features and might therefore miss crucial neurobiological memory mechanisms.

Storage in the mental lexicon requires persistent adaptation on longer timescales than unification. Engrams in long-term memory are viewed as strongly connected cell assemblies that encode information into synaptic conductances through STDP (Caporale and Dan, 2008; Miehl et al., 2022; Poo et al., 2016, but see Gallistel, 2021). There is less consensus, however, on whether engrams are exclusively located in excitatory synapses or also involve inhibitory ones (Hennequin et al., 2017), perhaps even primarily (Mongillo et al., 2018). Previous work has shown that engrams can emerge from sparse, random networks when multiple mechanisms for

unsupervised learning and homeostatic regulation interact dynamically (Litwin-Kumar & Doiron, 2014; Zenke et al., 2015). In these simulations, acquired memories were relatively stable in the presence of background noise and ongoing plasticity, and could be reactivated reliably after delay. These causal models of long-term storage can serve as a starting point for a neurobiological model of the mental lexicon. A promising first step in this direction has been taken in Tomasello et al. (2018).

Words in the mental lexicon have a feature structure consisting of, among others, semantic, syntactic and morphological attributes that are stored and maintained in the neurobiological infrastructure of the brain.⁵ In retrieval, speech sounds or letter sequences have to be recognized as particular words while, concurrently, these features are being computed from partial cues (pattern completion). Hence, there are at least two computational tasks that need to be solved in lexical retrieval; they happen in parallel and are likely to interact. For example, word recognition itself might sharpen the selection of features activated prior to the recognition point, perhaps through lateral inhibition. The computation of lexical features is currently not addressed by existing models that have focused on recognition only (e.g., those reviewed in Hannagan et al., 2013; Magnuson et al., 2020; Weber & Scharenborg, 2012). A causal model of the mental lexicon is needed that can explain how words are represented within the neurobiological substrate and how their feature structure is “activated” from perceptual input (Poeppel & Idsardi, 2022). Furthermore, the mental lexicon is language-specific, rapidly acquired in development through local learning mechanisms, and uniquely human. To explain these traits in neurobiological terms is another important challenge for causal modeling (see Open Questions).

The mental lexicon and unification continuously interact through feedback loops and exert reciprocal control (Figure 5). The feature structure of retrieved words controls the combinatorial operations of the unification network and, conversely, the partial interpretations computed by unification control the context-dependent retrieval process when multiple candidates are

⁵ For brevity, we refer to the content of the mental lexicon as ‘words’ which does not exclude larger units such as, e.g., collocations, multi-word expressions, idioms, or argument-structure constructions.

compatible with the sensory signal. To develop a combined architecture for adult language processing, the synaptic pathways for information exchange between these different functional modules can be fine-tuned using methods from control theory (Kao & Hennequin, 2019), feedback learning (Nicola & Clopath, 2017), or error-based optimization of networks (Neftci et al., 2019).

Compared to other cognitive domains, causal language modeling is in a privileged position because linguistic theory/analysis provide an extensive list of conceptual primitives that form the elementary units of language (which has been referred to as the ‘parts list’, Poeppel, 2012). In addition, a basic functional architecture can be derived from findings in cognitive neuroscience and the theory of computation (Figure 5). Hence, causal language modeling can draw on a rich set of reference points across Marr’s descriptive hierarchy; we know, roughly, which units and procedures to look for in neurobiology. However, if conceptual primitives and computational routines cannot be explicated in neurobiological terms, their theoretical status may eventually have to be revised.

Models of behavior versus the system

Computational language models that operate at the algorithmic level are often tested against linguistic behavior, i.e., system output or data collected in some experiment. The better a model reproduces or predicts behavior, the better it is considered to be validated. There are other adequacy criteria as well, but behavioral fit is a primary source of evidence in cognitive modeling. Causal models, on the other hand, are mainly concerned with the neurobiological mechanisms of the underlying system. They aim to be explanatory at the level of implementational causality: how do inputs give rise to outputs within the neurobiological machinery for language? Causal models are therefore not primarily about observations or behavior but—in the first instance—about the mechanisms that generate behavior. The extent to which a causal model behaves human-like is determined by the degree to which it approximates the biophysical characteristics of the actual system; fit with behavioral data is an independent outcome and not the immediate modeling goal (Box 2). This approach is reminiscent of early connectionist

language models which also intended to derive behavior from principles of neural information processing (e.g., Elman, 1990; McClelland et al., 1989). Today, it is widely held that these models incorporate too little neurobiological detail (see Figure 2) to be viewed as causal models of the neurobiological system (Arbib et al., 2000; Craver, 2006; Karaminis & Thomas, 2012). Deep learning approaches to natural language processing (see Young et al., 2018, for an overview), which are an extension of the connectionist paradigm, are very powerful in generating language-like output and might be a useful heuristic. However, Large Language Models (LLMs) are neither models of human behavior nor models of the neurobiological machinery. They do not model the causal structure of the language system nor cognitive function as such (language comprehension differs from next-word prediction, Bender & Koller, 2020), and they are sometimes inadequate behaviorally in that they fail in non-human ways and do not fail in human ways (Marcus, 2018, but see Linzen and Baroni, 2021, for a different perspective). Using LLMs to fit brain data (e.g., Goldstein et al., 2022; Schrimpf et al., 2021) is correlational rather than causal in nature. Hence, it is debatable whether they contribute novel insights to the study of human language at the implementational level of Marr's hierarchy.

In models of behavior, variables and parameters are dimensionless scalars that do not correspond to measurable quantities in biological reality and often lack interpretability in cognitive terms (Eckstein et al., 2022). In causal models, they have physical units of measurement (e.g., mV, nS, pF, etc.) that need to fall within physiological bounds. This restricts parameter choices to empirical ranges, reduces degrees of freedom, and puts strong constraints on the model space M (see Figure 3). Since units have to match on both sides of dynamical equations, causal models are also internally consistent. Whereas cognitive models often attempt to capture behavior with as few parameters as possible, the challenge for causal modeling is to deal with the abundance of parameters provided by the neurobiological system (e.g., on the order of $\approx 10^{14}$ synaptic conductances).⁶ Consequently, standard model selection criteria do not apply in causal modeling (e.g., Occam's razor). What needs to be explained is how the neurobiological language

⁶ Geoffrey Hinton refers to this distinction as the statistician's versus the neuroscientific perspective.

system can generalize appropriately despite being nominally overparameterized (Hasson et al., 2020). A third difference concerns the relationship between model time and real physical time. In cognitive models of behavior, time is often expressed in terms of processing steps and the relation to physical time is typically arbitrary. In causal models, time corresponds to real physical time since it arises from the dynamics of neuronal integration and synaptic transmission (Gerstner et al., 2014). Due to this inherent correspondence, a causal model would allow us, in principle, to investigate how speech and language processing unfold in time at any desired resolution. More importantly, however, causal models are therefore strongly constrained by real-time processing requirements whereas models of behavior typically are not.

Another difference between models of behavior and causal models of the system is related to their explanatory status. Output or behavior of a system should not be mistaken for the mechanisms that generate behavior at the level of physical, or neurobiological, causes. For example, a statistical model of weather data can have high predictive accuracy but it is not a model of Earth's atmosphere that generates the weather.⁷ By parity of reasoning, suppose a cognitive language model reproduces all known behavioral data. This would not guarantee that the model correctly describes the algorithms employed by the brain and it would still be unclear whether the model is explanatory with respect to the causal generators of behavior. This uncertainty persists until it has been demonstrated that a proposed algorithmic model can be reduced to the relevant neurobiology. Similar uncertainty afflicts experimental approaches that attempt to reverse engineer the computational machinery for language from behavioral output. Neuroimaging methods (fMRI, EEG, MEG, etc.) observe sequences of brain states, i.e., processing outcomes or system behavior in the broadest sense, but not the neurobiological processing dynamics itself which is hidden from the measurement devices. Another complication is that the fMRI signal, for example, is related to the BOLD response which in turn is related indirectly to neural activity. Language models that are inferred from such data are confounded by

⁷ Likewise, no one would confuse a regression model of experimental data with a model of the processes that generated the data.

these theoretical linking principles which need to be factored out in order to arrive at a veridical model of the neurobiological processing machinery.

Simulations of the language system at the level of implementational causality are not confounded in this way and enable us to study candidate processing dynamics with unrivaled spatio-temporal precision. Moreover, component parts that lack neurobiological support do not enter into model design to begin with. Reduction has already been achieved at the level of computational elements and their interaction. Hence, neurobiological causal models describe the mechanistic generators of linguistic behavior from which observed behavior can be derived. Without a neurobiological foundation, modeling behavior is not explanatory with respect to the causal generators of behavior unless such models can be shown to be reducible to neurobiology. It is understood that a causal modeling approach requires a long-term perspective; it will take time and effort for it to succeed.

Causality, reduction and abstraction

David Marr considered the functional level to be the most important one for understanding biological information processing systems but emphasized that different questions need to be addressed at different levels of description. He also pointed out that the different levels are “logically and causally related” (Marr, 1982, p. 25). In particular, the algorithmic level is not autonomous with respect to the implementational level. Amongst a number of candidate language models, it is neurobiology that is going to select the correct one, if any. There might be multiple abstractions that are equivalent in some deep sense but there is still a matter of fact in the brain which of these abstractions is valid. For instance, recursive function theory itself can be formulated within many different mathematical frameworks, but it is an empirical question which algorithmic model the brain implements to “run” this theory.⁸ Thus, although methodologically any of Marr’s levels can serve as a starting point, neurobiology is *ontologically prior* since it determines the algorithms that are implemented by the real system which, ultimately, also

⁸ Surely, the brain does not implement language as Conway’s *Game of Life* or *Baba Is You*, both of which are Turing universal.

determine the range of possible language behavior we can observe. Both algorithm and behavior are caused by the underlying neurobiology, while the converse is not true.

In light of these dependencies, we should therefore not be satisfied to describe language at a single level only; the ambition must be to link and traverse levels through explanatory bridging principles. As an analogy, a structured computer architecture with its many layers of abstraction can be used through an operating system because the interfaces between layers are correctly designed (Tanenbaum & Austin, 2013). In other words, a higher level of abstraction has to comply with and systematically relate to lower level mechanisms by reduction. Thus, it is only under the condition of reducibility that we can “ignore” lower levels. What is currently underspecified in cognitive theories of language are precisely these interfaces between levels of abstraction. Despite decades of computational work it has not been possible to connect cognitive language models to neurobiology in a substantial manner. With a few notable exceptions (Fitz et al., 2020; Rolls & Deco, 2015; Tomasello et al., 2018), models of language processing that are characterized as ‘neurocomputational’ or ‘neurally plausible’ do not yet make sufficient contact with the basic neurobiological principles described in Figure 2. This also holds for language models in deep learning. The assumption that we can abstract away from these principles needs to be scientifically justified because abstraction without reduction is likely to result in simplifications that may not be valid.

Within the computer metaphor, the terminology of cognitive theory is comparable to a high-level programming language, like Python or Julia. Underneath this layer of abstraction lies the hardware-dependent ‘machine language’ of the implementational substrate. The machine language determines the basic set of instructions, data types and memory registers that are instantiated by the actual neurobiological system. This ‘instruction set architecture’ (ISA) corresponds to circuits built from biological neurons, their membrane potentials, spike generation mechanisms, synaptic currents, dendritic integration, etc. A cognitive theory of language that is empirically adequate must be realizable in this neurobiological ISA, otherwise it remains disconnected from the implementational level of description. Causal models, on the other hand,

are directly formulated in the language of the neurobiological ISA and pinpoint the fundamental computational elements in neurobiology, their interactions, and how they support language functions.

Although causal models describe language processing in terms of biological neurons and synapses, one long-term goal is to abstract a homomorphic cognitive model from the neurobiological specification that instantiates a correct algorithmic description of the language faculty. There is, of course, no guarantee that any *particular* causal model will yield a correct cognitive theory. But any cognitive theory that is correct needs to be consistent with what is known about the language system from a neurobiological perspective.⁹ Through simulation, analysis and theoretical insight, the aim is to *discover* rather than guess the algorithms that operate at the neural level. These algorithms, in addition, have to explain the breathtaking speed, fault tolerance, and energy efficiency of the brain system for language. The functionalist doctrine and multiple realizability, which are only concerned with non-biological input-output relations, have no bearing on these issues.

Validation of causal models involves different sources of evidence, including behavior, none of which is sufficient on its own. In this sense, causal modeling is not intrinsically reductionist but aims to encompass all of Marr's levels in the final analysis (Figure 1 and Box 2). Models that are behaviorally adequate but violate known neurobiology cannot be correct. Models that are behaviorally inadequate but consistent with known neurobiology need to be refined. Thus, causal modeling advocates an iterative approach that seeks to gradually approximate language behavior from first principles of neurobiology, through cycles of model development, validation and revision.

Concluding remarks

The nature of the language faculty—its representations, storage mechanisms, and elementary operations—is determined by the neurobiological infrastructure that sustains it. A large number of replicable findings from experimental neuroscience (Luo, 2015; Sterling &

⁹ A similar point has been made by Feldman (2006) in more general terms.

Laughlin, 2015) have been formalized as effective mathematical models (Gerstner et al., 2014) that can readily be used as the basic building blocks for causal language modeling. Complex systems assembled from these neurobiological component parts are analytically intractable, and simulation therefore becomes a methodological necessity (Einevoll et al., 2019; Gerstner et al., 2012). With unprecedented access to computational power and neurobiological insight to constrain these simulations, it is the appropriate time to supplement traditional methods in language research with causal modeling in order to integrate language across levels of description. Neurobiological causal modeling follows the classical path of science in attempting to understand complex systems—e.g., multicellular organisms, condensed matter, or planetary climate—from observations, to statistical modeling, to explaining the causal structure of the physical system that generated the observations in the first place. Eventually, this approach might even allow us to bridge into the genetic basis of language.

Computational models of cognitive function are in need of stronger neurobiological foundations (O'Reilly, 2006) and several recent perspective articles have similarly suggested to “close the mechanistic gap” by means of neurobiologically-grounded models of information processing (Paquola et al., 2022; Pulvermüller et al., 2021). Our proposal is focusing on the language domain where computational models have played a particularly prominent role. However, neurobiological causal modeling amounts to more than neural network modeling with a few added constraints. Rather, we propose to reconceptualize computational language modeling and start building causal models from the ground up. This approach will not only address the missing interfaces between levels of description but is also expected to have profound ramifications at the algorithmic and functional levels themselves (Larkum, 2022). We call to action the community of language researchers to engage with this complementary approach and confront the challenges of investigating the neurobiology of language on the basis of first principles of brain organization. A joint, multidisciplinary effort is needed to bring this research program to fruition.

Causal models are formulated as systems of coupled differential equations which is the

lingua franca of science. They describe the fundamental dynamical principles underlying cognitive function in neurobiology. Hence, they provide a common, unified framework for modeling cognition that makes different instantiations of causal models commensurable and falsifiable (Haeffel, 2022; Popper, 1959). In the long term, this approach will lead to better theories of language processing, the progressive accumulation of scientific knowledge, and a deeper understanding not only of language but other cognitive phenomena as well.

Funding information

This work was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

- Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511574566>
- Arbib, M. A., Billard, A., Iacoboni, M., & Oztop, E. (2000). Synthetic brain imaging: Grasping, mirror neurons and imitation. *Neural Networks*, 13, 975–997. [https://doi.org/10.1016/S0893-6080\(00\)00070-8](https://doi.org/10.1016/S0893-6080(00)00070-8)
- Astle, D. E., Johnson, M. H., & Akarca, D. (2023). Toward computational neuroconstructivism: A framework for developmental systems neuroscience. *Trends in Cognitive Sciences*, 27(8), 726–744. <https://doi.org/10.1016/j.tics.2023.04.009>
- Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21(10), 1133–1145. <https://doi.org/10.1097/00004647-200110000-00001>
- Barak, O., & Tsodyks, M. (2014). Working models of working memory. *Current Opinion in Neurobiology*, 25, 20–24. <https://doi.org/10.1016/j.conb.2013.10.008>
- Bargmann, C. I. (2012). Beyond the connectome: How neuromodulators shape neural circuits. *Bioessays*, 34, 458–465. <https://doi.org/10.1002/bies.201100185>
- Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55, 55–64. <https://doi.org/10.1016/j.conb.2019.01.007>
- Bartol Jr, T. M., Bromer, C., Kinney, J., Chirillo, M. A., Bourne, J. N., Harris, K. M., & Sejnowski, T. J. (2015). Nanoconnectomic upper bound on the variability of synaptic plasticity. *eLife*, 4, 1–18. <https://doi.org/10.7554/eLife.10778>
- Bechtel, W. (2022). Reductionistic explanations of cognitive information processing: Bottoming out in neurochemistry. *Frontiers in Integrative Neuroscience*, 16, 944303. <https://doi.org/10.3389/fnint.2022.944303>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the*

- Association for Computational Linguistics*, 5185–5198.
<https://doi.org/10.18653/v1/2020.acl-main.463>
- Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1), 65–98. <https://doi.org/10.1137/141000671>
- Binzegger, T., Douglas, R. J., & Martin, K. A. (2004). A quantitative map of the circuit of cat primary visual cortex. *Journal of Neuroscience*, 24, 8441–8453.
<https://doi.org/10.1523/JNEUROSCI.1400-04.2004>
- Bonaiuto, J., & Arbib, M. A. (2014). Modeling the BOLD correlates of competitive neural dynamics. *Neural Networks*, 49, 1–10. <https://doi.org/10.1016/j.neunet.2013.09.001>
- Bray, D. (2009). *Wetware: A computer in every living cell*. Yale University Press.
- Brette, R. (2015). Philosophy of the spike: Rate-based vs. spike-based theories of the brain. *Frontiers in Systems Neuroscience*, 9, 151. <https://doi.org/10.3389/fnsys.2015.00151>
- Brette, R., & Gerstner, W. (2005). Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. *Journal of Neurophysiology*, 94, 3637–3642.
<https://doi.org/10.1152/jn.00686.2005>
- Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, 10, 113–125.
<https://doi.org/10.1038/nrn2558>
- Buzsáki, G. (2010). Neural syntax: Cell assemblies, synapse ensembles, and readers. *Neuron*, 68(3), 362–385. <https://doi.org/10.1016/j.neuron.2010.09.023>
- Caporale, N., & Dan, Y. (2008). Spike timing–dependent plasticity: A Hebbian learning rule. *Annual Review of Neuroscience*, 31(1), 25–46.
<https://doi.org/10.1146/annurev.neuro.31.060407.125639>
- Cavallari, S., Panzeri, S., & Mazzoni, A. (2014). Comparison of the dynamics of neural interactions between current-based and conductance-based integrate-and-fire recurrent networks. *Frontiers in Neural Circuits*, 8, 12. <https://doi.org/10.3389/fncir.2014.00012>

- Chaudhuri, R., & Fiete, I. (2016). Computational principles of memory. *Nature Neuroscience*, 19, 394–403. <https://doi.org/10.1038/nn.4237>
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.
- Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science*, 242(4879), 741–745. <https://doi.org/10.1126/science.3055294>
- Churchland, P. S., & Sejnowski, T. J. (Eds.). (1992). *The computational brain*. MIT Press.
- Clopath, C. (2012). Synaptic consolidation: An approach to long-term learning. *Cognitive Neurodynamics*, 6(3), 251–257. <https://doi.org/10.1007/s11571-011-9177-6>
- Clopath, C., Büsing, L., Vasilaki, E., & Gerstner, W. (2010). Connectivity reflects coding: A model of voltage-based STDP with homeostasis. *Nature Neuroscience*, 11, 344–352. <https://doi.org/10.1038/nn.2479>
- Clopath, C., Ziegler, L., Vasilaki, E., Büsing, L., & Gerstner, W. (2008). Tag-trigger-consolidation: A model of early and late long-term-potential and depression. *PLoS Computational Biology*, 4(12), e1000248. <https://doi.org/10.1371/journal.pcbi.1000248>
- Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153, 355–376. <https://doi.org/10.1007/s11229-006-9097-x>
- Cutland, N. (1980). *Computability: An introduction to recursive function theory*. Cambridge University Press.
- Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., & Pallier, C. (2015). The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88, 2–19. <https://doi.org/10.1016/j.neuron.2015.09.019>
- Destexhe, A., Mainen, Z. F., & Sejnowski, T. J. (1998). Kinetic models of synaptic transmission. In C. Koch & I. Segev (Eds.), *Methods in neuronal modeling* (2nd edition, pp. 1–25, Vol. 2). MIT Press.

- Ding, Y., Wang, Y., & Cao, L. (2022). A simplified plasticity model based on synaptic tagging and capture theory: Simplified STC. *Frontiers in Computational Neuroscience*, 15, 798418. <https://doi.org/10.3389/fncom.2021.798418>
- Duarte, R., & Morrison, A. (2019). Leveraging heterogeneity for neural computation with fading memory in layer 2/3 cortical microcircuits. *PLoS Computational Biology*, 15(4), e1006781. <https://doi.org/10.1371/journal.pcbi.1006781>
- Duarte, R., Seeholzer, A., Zilles, K., & Morrison, A. (2017). Synaptic patterning and the timescales of cortical dynamics. *Current Opinion in Neurobiology*, 43, 156–165. <https://doi.org/10.1016/j.conb.2017.02.007>
- Duarte, R., Uhlmann, M., van den Broek, D., Fitz, H., Petersson, K. M., & Morrison, A. (2018). Encoding symbolic sequences with spiking neural reservoirs. *Proceedings of the International Joint Conference on Neural Networks*, 1–8. <https://doi.org/10.1109/IJCNN.2018.8489114>
- Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, 3, 1184–1191. <https://doi.org/10.1038/81460>
- Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2022). The interpretation of computational model parameters depends on the context. *eLife*, 11, e75474. <https://doi.org/10.7554/eLife.75474>
- Einevoll, G. T., Destexhe, A., Diesmann, M., Grün, S., Jirsa, V., de Kamps, M., Migliore, M., Ness, T. V., Plesser, H. E., & Schürmann, F. (2019). The scientific case for brain simulations. *Neuron*, 102, 735–744. <https://doi.org/10.1016/j.neuron.2019.03.027>
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32, 429–492. <https://doi.org/10.1017/S0140525X0999094X>

- Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, 9, 292–303. <https://doi.org/10.1038/nrn2258>
- Feldman, J. A. (2006). *From molecule to metaphor—a neural theory of language*. MIT Press.
- Fiebig, F., & Lansner, A. (2017). A spiking working memory model based on Hebbian short-term potentiation. *Journal of Neuroscience*, 37(1), 83–96. <https://doi.org/10.1523/jneurosci.1989-16.2016>
- Fields, R. D. (2022). The enigma of working memory: Changing views. *Neuroscientist*, 28(5), 420–424. <https://doi.org/10.1177/10738584211072747>
- Fişek, M., & Häusser, M. (2020). Are human dendrites different? *Trends in Cognitive Sciences*, 24(6), 411–412. <https://doi.org/10.1016/j.tics.2020.03.002>
- Fitz, H., Uhlmann, M., van den Broek, D., Duarte, R., Hagoort, P., & Petersson, K. M. (2020). Neuronal spike-rate adaptation supports working memory in language processing. *Proceedings of the National Academy of Sciences of the U.S.A.*, 117(34), 20881–20889. <https://doi.org/10.1073/pnas.2000222117>
- Frémaux, N., & Gerstner, W. (2016). Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in Neural Circuits*, 9, 85. <https://doi.org/10.3389/fncir.2015.00085>
- Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modeling. *NeuroImage*, 19, 1273–1302. [https://doi.org/10.1016/S1053-8119\(03\)00202-7](https://doi.org/10.1016/S1053-8119(03)00202-7)
- Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science*, 173, 652–654. <https://doi.org/10.1126/science.173.3997.652>
- Gallistel, C. R. (2021). The physical basis of memory. *Cognition*, 213, 104533. <https://doi.org/10.1016/j.cognition.2020.104533>
- Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition*. Cambridge University Press.

- Gerstner, W., Kreiter, A. K., Markram, H., & Herz, A. V. M. (1997). Neural codes: Firing rates and beyond. *Proceedings of the National Academy of Sciences of the U.S.A.*, 94, 12740–12741. <https://doi.org/10.1073/pnas.94.24.12740>
- Gerstner, W., Sprekeler, H., & Deco, G. (2012). Theory and simulation in neuroscience. *Science*, 338, 60–65. <https://doi.org/10.1126/science.1227356>
- Gewaltig, M.-O., & Diesmann, M. (2007). NEST (NEural Simulation Tool). *Scholarpedia*, 2(4), 1430.
- Gidon, A., Zolnik, T. A., Fidzinski, P., Bolduan, F., Papoutsi, A., Poirazi, P., Holtkamp, M., Vida, I., & Larkum, M. E. (2020). Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, 367(6473), 83–87. <https://doi.org/10.1126/science.aax6239>
- Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, 14(3), 477–485. [https://doi.org/10.1016/0896-6273\(95\)90304-6](https://doi.org/10.1016/0896-6273(95)90304-6)
- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., ... Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369–380. <https://doi.org/10.1038/s41593-022-01026-4>
- Gouwens, N. W., Sorensen, S. A., Berg, J., Lee, C., Jarsky, T., Ting, J., Sunkin, S. M., Feng, D., Anastassiou, C. A., Barkan, E., Bickley, K., Blesie, N., Braun, T., Brouner, K., Budzillo, A., Caldejon, S., Casper, T., Castelli, D., Chong, P., ... Koch, C. (2019). Classification of electrophysiological and morphological neuron types in the mouse visual cortex. *Nature Neuroscience*, 22, 1182–1195. <https://doi.org/10.1038/s41593-019-0417-0>
- Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424–438. <https://doi.org/10.2307/1912791>

- Groschner, L. N., Malis, J. G., Zuidinga, B., & Borst, A. (2022). A biophysical account of multiplication by a single neuron. *Nature*, 603, 119–123.
<https://doi.org/10.1038/s41586-022-04428-3>
- Haber, S. N. (2016). Corticostriatal circuitry. *Dialogues in Clinical Neuroscience*, 18(1), 7–21.
<https://doi.org/10.31887/DCNS.2016.18.1/shaber>
- Haefffel, G. J. (2022). Psychology needs to get tired of winning. *Royal Society Open Science*, 9, 220099. <https://doi.org/10.1098/rsos.220099>
- Hagen, E., Dahmen, D., Stavrinou, M. L., Lindén, H., Tetzlaff, T., van Albada, S. J., Grün, S., Diesmann, M., & Einevoll, G. T. (2016). Hybrid scheme for modeling local field potentials from point-neuron networks. *Cerebral Cortex*, 26(12), 4461–4496.
<https://doi.org/10.1093/cercor/bhw237>
- Hagoort, P. (2005). On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, 9(9), 416–423. <https://doi.org/10.1016/j.tics.2005.07.004>
- Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, 366, 55–58. <https://doi.org/10.1126/science.aax0289>
- Hannagan, T., Magnuson, J., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, 4, 563. <https://doi.org/10.3389/fpsyg.2013.00563>
- Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in Cognitive Sciences*, 19(6), 304–313. <https://doi.org/10.1016/j.tics.2015.04.006>
- Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3), 416–434.
<https://doi.org/10.1016/j.neuron.2019.12.002>
- Hennequin, G., Agnes, E., & Vogels, T. (2017). Inhibitory plasticity: Balance, control, and codependence. *Annual Review of Neuroscience*, 40(1), 557–579.
<https://doi.org/10.1146/annurev-neuro-072116-031005>

- Herstel, L. J., & Wieringa, C. J. (2021). Network control through coordinated inhibition. *Current Opinion in Neurobiology*, 67, 34–41. <https://doi.org/10.1016/j.conb.2020.08.001>
- Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.
- Jackendoff, R., & Audring, J. (2020). *The texture of the lexicon: Relational morphology and the parallel architecture*. Oxford University Press.
- Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLoS Computational Biology*, 13(1), e1005268. <https://doi.org/10.1371/journal.pcbi.1005268>
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2012). *Principles of neural science*. McGraw-Hill.
- Kao, T.-C., & Hennequin, G. (2019). Neuroscience out of control: Control-theoretic perspectives on neural circuit dynamics. *Current Opinion in Neurobiology*, 58, 122–129. <https://doi.org/10.1016/j.conb.2019.09.001>
- Karaminis, T. N., & Thomas, M. S. C. (2012). Connectionism. In N. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 767–771). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_397
- Kass, R. E., Amari, S. I., Arai, K., Brown, E. N., Diekman, C. O., Diesmann, M., Doiron, B., Eden, U. T., Fairhall, A. L., Fiddymment, G. M., Fukai, T., Grün, S., Harrison, M. T., Helias, M., Nakahara, H., Teramae, J. N., Thomas, P. J., Reimers, M., Rodu, J., . . . Kramer, M. A. (2018). Computational neuroscience: Mathematical and statistical perspectives. *Annual Review of Statistics and Its Application*, 5, 183–214. <https://doi.org/10.1146/annurev-statistics-041715-033733>
- Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. Oxford University Press.
- Lafourcade, M., van der Goes, M.-S. H., Vardalaki, D., Brown, N. J., Voigts, J., Yun, D.-H., Kim, M. E., Ku, T., & Harnett, M. T. (2022). Differential dendritic integration of long-range inputs in association cortex via subcellular changes in synaptic

- AMPA-to-NMDA receptor ratio. *Neuron*, 110, 1532–1546.e4.
<https://doi.org/10.1016/j.neuron.2022.01.025>
- Larkum, M. E. (2022). Are dendrites conceptually useful? *Neuroscience*, 489, 4–14.
<https://doi.org/10.1016/j.neuroscience.2022.03.008>
- Levy, S., & Bargmann, C. I. (2020). An adaptive-threshold mechanism for odor sensation and animal navigation. *Neuron*, 105(3), 534–548.e13.
<https://doi.org/10.1016/j.neuron.2019.10.034>
- Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7, 195–212. <https://doi.org/10.1146/annurev-linguistics-032020-051035>
- Lisman, J. E., Fellous, J.-M., & Wang, X.-J. (1998). A role for NMDA-receptor channels in working memory. *Nature Neuroscience*, 1(4), 273–275. <https://doi.org/10.1038/1086>
- Litwin-Kumar, A., & Doiron, B. (2014). Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nature Communications*, 5, 5319.
<https://doi.org/10.1038/ncomms6319>
- Loewenstein, Y., & Sompolinsky, H. (2003). Temporal integration by calcium dynamics in a model neuron. *Nature Neuroscience*, 6, 961–967. <https://doi.org/doi.org/10.1038/nn1109>
- London, M., & Häusser, M. (2005). Dendritic computation. *Annual Review of Neuroscience*, 28, 503–532. <https://doi.org/10.1146/annurev.neuro.28.061604.135703>
- Luo, L. (2015). *Principles of neurobiology*. Garland Science.
- Luz, Y., & Shamir, M. (2012). Balancing feed-forward excitation and inhibition via Hebbian inhibitory synaptic plasticity. *PLoS Computational Biology*, 8(1), e1002334.
<https://doi.org/10.1371/journal.pcbi.1002334>
- Maass, W., & Orponen, P. (1998). On the effect of analog noise in discrete-time analog computations. *Neural Computation*, 10(5), 1071–1095.
<https://doi.org/10.1162/089976698300017359>
- Magee, J. C., & Grienberger, C. (2020). Synaptic plasticity forms and functions. *Annual Review of Neuroscience*, 43(1), 95–117. <https://doi.org/10.1146/annurev-neuro-090919-022842>

- Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, 44, e12823. <https://doi.org/10.1111/cogs.12823>
- Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv*, 1801, 00631. <https://doi.org/10.48550/arXiv.1801.00631>
- Marder, E., Abbott, L. F., Turrigiano, G. G., Liu, Z., & Golowasch, J. (1996). Memory from the dynamics of intrinsic membrane currents. *Proceedings of the National Academy of Sciences of the U.S.A.*, 93(24), 13481–13486. <https://doi.org/10.1073/pnas.93.24.13481>
- Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297), 213–215. <https://doi.org/10.1126/science.275.5297.213>
- Markram, H., Wang, Y., & Tsodyks, M. (1998). Differential signaling via the same axon of neocortical pyramidal neurons. *Proceedings of the National Academy of Sciences of the U.S.A.*, 95(9), 5323–5328. <https://doi.org/10.1073/pnas.95.9.5323>
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Freeman.
- Martin, A. E., & Doumas, L. A. A. (2017). A mechanism for the cortical computation of hierarchical linguistic structure. *PLoS Biology*, 15(3), e2000663. <https://doi.org/10.1371/journal.pbio.2000663>
- Mazzoni, A., Lindén, H., Cuntz, H., Lansner, A., Panzeri, S., & Einevoll, G. T. (2015). Computing the local field potential (LFP) from integrate-and-fire network models. *PloS Computational Biology*, 11, e1004584. <https://doi.org/10.1371/journal.pcbi.1004584>
- McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4, SI 287–335. <https://doi.org/10.1080/01690968908406371>

- McDougal, R. A., Morse, T. M., Carnevale, T., Marengo, L., Wang, R., Migliore, M., Miller, P. L., Shepherd, G. M., & Hines, M. L. (2017). Twenty years of ModelDB and beyond: Building essential modeling tools for the future of neuroscience. *Journal of Computational Neuroscience*, 42(1), 1–10. <https://doi.org/10.1007/s10827-016-0623-7>
- Miehl, C., Onasch, S., Festa, D., & Gjorgjieva, J. (2022). Formation and computational implications of assemblies in neural circuits. *Journal of Physiology*. <https://doi.org/10.1113/JP282750>
- Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, 319, 1543–1546. <https://doi.org/10.1126/science.1150769>
- Mongillo, G., Rumpel, S., & Loewenstein, Y. (2018). Inhibitory connectivity defines the realm of excitatory plasticity. *Nature Neuroscience*, 21(10), 1463–1470. <https://doi.org/10.1038/s41593-018-0226-x>
- Mountcastle, V. B. (1997). The columnar organization of the neocortex. *Brain*, 120, 701–722. <https://doi.org/10.1093/brain/120.4.701>
- Neftci, E. O., Mostafa, H., & Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6), 51–63. <https://doi.org/10.1109/MSP.2019.2931595>
- Nicola, W., & Clopath, C. (2017). Supervised learning in spiking neural networks with FORCE training. *Nature Communications*, 8, 2208. <https://doi.org/10.1038/s41467-017-01827-3>
- Nolte, M., Reimann, M. W., King, J. G., Markram, H., & Muller, E. B. (2019). Cortical reliability amid noise and chaos. *Nature Communications*, 10, 3792. <https://doi.org/10.1038/s41467-019-11633-8>
- O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition. *Science*, 314, 91–94. <https://doi.org/10.1126/science.1127242>

- Papadimitriou, C. H., & Friederici, A. D. (2022). Bridging the gap between neurons and cognition through assemblies of neurons. *Neural Computation*, 34, 291–306.
https://doi.org/10.1162/neco_a_01463
- Papoutsis, A., Sidiropoulou, K., & Poirazi, P. (2014). Dendritic nonlinearities reduce network size requirements and mediate ON and OFF states of persistent activity in a PFC microcircuit model. *PLoS Computational Biology*, 10, e1003764.
<https://doi.org/10.1371/journal.pcbi.1003764>
- Paquola, C., Amunts, K., Evans, A., Smallwood, J., & Bernhardt, B. (2022). Closing the mechanistic gap: The value of microarchitecture in understanding cognitive networks. *Trends in Cognitive Sciences*, 26, 873–886. <https://doi.org/10.1016/j.tics.2022.07.001>
- Payeur, A., Béique, J.-C., & Naud, R. (2019). Classes of dendritic information processing. *Current Opinion in Neurobiology*, 58, 78–85. <https://doi.org/10.1016/j.conb.2019.07.006>
- Payeur, A., Guerguiev, J., Zenke, F., Richards, B. A., & Naud, R. (2021). Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature Neuroscience*, 24(7), 1010–1019. <https://doi.org/10.1038/s41593-021-00857-x>
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.
- Petersson, K. M. (2005). On the relevance of the neurobiological analogue of the finite-state architecture. *Neurocomputing*, 65(66), 825–832.
<https://doi.org/doi:10.1016/j.neucom.2004.10.108>
- Petersson, K. M. (2008). On cognition, structured sequence processing, and adaptive dynamical systems. *AIP Conference Proceedings*, 1060, 195–200. <https://doi.org/10.1063/1.3037051>
- Petersson, K. M., & Hagoort, P. (2012). The neurobiology of syntax: Beyond string-sets. *Philosophical Transactions of the Royal Society B*, 367(1598), 1971–1883.
<https://doi.org/10.1098/rstb.2012.0101>
- Petrides, M. (2014). *Neuroanatomy of language regions of the human brain*. Academic Press.

- Pfister, J.-P., & Gerstner, W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. *Journal of Neuroscience*, 26, 9673–9682.
<https://doi.org/10.1523/jneurosci.1425-06.2006>
- Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, 29(1–2), 34–55.
<https://doi.org/10.1080/02643294.2012.710600>
- Poeppel, D., & Idsardi, W. (2022). We don't know how the brain stores anything, let alone words. *Trends in Cognitive Sciences*, 26, 1054–1055. <https://doi.org/10.1016/j.tics.2022.08.010>
- Poirazi, P., & Papoutsis, A. (2020). Illuminating dendritic function with computational models. *Nature Review Neuroscience*, 21(6), 303–321. <https://doi.org/10.1038/s41583-020-0301-7>
- Poo, M.-M., Pignatelli, M., Ryan, T. J., Tonegawa, S., Bonhoeffer, T., Martin, K. C., Rudenko, A., Tsai, L.-H., Tsien, R. W., Fishell, G., Mullins, C., Gonçalves, J. T., Shtrahman, M., Johnston, S. T., Gage, F. H., Dan, Y., Long, J., Buzsáki, G., & Stevens, C. (2016). What is memory? The present state of the engram. *BMC Biology*, 14, 40.
<https://doi.org/10.1186/s12915-016-0261-6>
- Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson of London.
- Pulvermüller, F., Tomasello, M. R., R. and Henningsen-Schomers, & Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, 22, 488–502. <https://doi.org/10.1038/s41583-021-00473-5>
- Quaresima, A., Fitz, H., Duarte, R., van den Broek, D., Hagoort, P., & Petersson, K. M. (2022). The Tripod neuron: A minimal structural reduction of the dendritic tree. *Journal of Physiology*, 601(15), 3265–3295. <https://doi.org/10.1113/JP283399>
- Rabinovich, M., Huerta, R., & Laurent, G. (2008). Transient dynamics for neural processing. *Science*, 321(5885), 48–50. <https://doi.org/10.1126/science.1155564>
- Rao, A., Plank, P., Wild, A., & Maass, W. (2022). A long short-term memory for AI applications in spike-based neuromorphic hardware. *Nature Machine Intelligence*, 4, 467–479.
<https://doi.org/10.1038/s42256-022-00480-w>

- Rigotti, M., Barak, O., Warden, M., Wang, X.-J., Daw, N., Miller, E., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451), 585–590. <https://doi.org/10.1038/nature12160>
- Rolls, E. T., & Deco, G. (2015). Networks for memory, perception, and decision-making, and beyond to how the syntax for language might be implemented in the brain. *Brain Research*, 1621, 316–334. <https://doi.org/10.1016/j.brainres.2014.09.021>
- Rossant, C., Goodman, D. F. M., Fontaine, B., Platkiewicz, J., Magnusson, A. K., & Brette, R. (2011). Fitting neuron models to spike trains. *Frontiers in Neuroscience*, 5(9), 1–8. <https://doi.org/10.3389/fnins.2011.00009>
- Roth, A., & van Rossum, M. (2009). Modeling synapses. In E. De Schutter (Ed.), *Computational modeling methods for neuroscientists* (pp. 139–159, Vol. 6). MIT Press.
- Roxin, A., Brunel, N., Hansel, D., Mongillo, G., & van Vreeswijk, C. (2011). On the distribution of firing rates in networks of cortical neurons. *Journal of Neuroscience*, 31(45), 16217–16226. <https://doi.org/10.1523/JNEUROSCI.1677-11.2011>
- Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, 55, 103–111. <https://doi.org/10.1016/j.conb.2019.02.002>
- Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the U.S.A.*, 118(45), e2105646118. <https://doi.org/10.1073/pnas.2105646118>
- Shepherd, G. M. (Ed.). (2004). *The synaptic organization of the brain* (5th). Oxford University Press.
- Siegelmann, H. T. (1999). *Neural networks and analog computation: Beyond the turing limit*. Birkhäuser.
- Softky, W. R., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience*, 13(1), 334–350. <https://doi.org/10.1523/JNEUROSCI.13-01-00334.1993>

- Spratling, M. W. (2002). Cortical region interactions and the functional role of apical dendrites. *Behavioral and Cognitive Neuroscience Reviews*, 1(3), 219–228.
<https://doi.org/10.1177/1534582302001003003>
- Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. MIT Press.
- Stimberg, M., Brette, R., & Goodman, D. (2019). Brian 2, an intuitive and efficient neural simulator (F. K. Skinner, Ed.). *eLife*, 8, e47314. <https://doi.org/10.7554/eLife.47314>
- Strata, P., & Harvey, R. (1999). Dale's principle. *Brain Research Bulletin*, 50(5), 349–350.
[https://doi.org/10.1016/s0361-9230\(99\)00100-8](https://doi.org/10.1016/s0361-9230(99)00100-8)
- Tanenbaum, A. S., & Austin, T. (2013). *Structured computer organization* (6th). Pearson.
- Tetzlaff, C., Kolodziejewski, C., Timme, M., & Wörgötter, F. (2012). Analysis of synaptic scaling in combination with Hebbian plasticity in several simple networks. *Frontiers in Computational Neuroscience*, 6, 36. <https://doi.org/10.3389/fncom.2012.00036>
- Titley, H. K., Brunel, N., & Hansel, C. (2017). Towards a neurocentric view of learning. *Neuron*, 95(1), 19–32. <https://doi.org/10.1016/j.neuron.2017.05.021>
- Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermüller, F. (2018). A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. *Frontiers in Computational Neuroscience*, 12, 88.
<https://doi.org/10.3389/fncom.2018.00088>
- Tremblay, P., & Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain and Language*, 162, 60–71.
<https://doi.org/10.1016/j.bandl.2016.08.004>
- Tripathy, S. J., Savitskaya, J., Burton, S. D., Urban, N. N., & Gerkin, R. C. (2014). Neuroelectro: A window to the world's neuron electrophysiology data. *Frontiers in Neuroinformatics*, 8, 40. <https://doi.org/10.3389/fninf.2014.00040>
- Turrigiano, G. G. (2008). The self-tuning neuron: Synaptic scaling of excitatory synapses. *Cell*, 135(3), 422–435. <https://doi.org/10.1016/j.cell.2008.10.008>

- Turrigiano, G. G., Marder, E., & Abbott, L. F. (1996). Cellular short-term memory from a slow potassium conductance. *Journal of Neurophysiology*, 75(2), 963–966.
<https://doi.org/10.1152/jn.1996.75.2.963>
- Turrigiano, G. G., & Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5, 97–107. <https://doi.org/10.1038/nrn1327>
- Uhlmann, M. (2020). *Neurobiological models of sentence processing* [Doctoral dissertation, Max Planck Institute for Psycholinguistics].
https://pure.mpg.de/rest/items/item_3251916_2/component/file_3261327/content
- van der Velde, F., & Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, 29, 37–70.
<https://doi.org/10.1017/S0140525X06009022>
- van Vreeswijk, C., & Sompolinsky, H. (1996). Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293), 1724–1726.
<https://doi.org/10.1126/science.274.5293.1724>
- Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C., & Gerstner, W. (2011). Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, 334(6062), 1569–1573. <https://doi.org/10.1126/science.1211095>
- Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3), 387–401.
<https://doi.org/10.1002/wcs.1178>
- Whittington, J. C. R., & Bogacz, R. (2019). Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, 23(3), 235–250. <https://doi.org/10.1016/j.tics.2018.12.005>
- Xue, M., Atallah, B. V., & Scanziani, M. (2014). Equalizing excitation–inhibition ratios across visual cortical neurons. *Nature*, 511, 596–600. <https://doi.org/10.1038/nature13321>
- Young, T., Hazarikaz, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55–75.
<https://doi.org/10.1109/MCI.2018.2840738>

- Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10, 3770.
<https://doi.org/10.1038/s41467-019-11786-6>
- Zenke, F., Agnes, E. J., & Gerstner, W. (2015). Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. *Nature Communications*, 6, 6922. <https://doi.org/10.1038/ncomms7922>
- Ziegler, L., Zenke, F., Kastner, D. B., & Gerstner, W. (2015). Synaptic consolidation: From synapses to behavioral modeling. *Journal of Neuroscience*, 35(3), 1319–1334.
<https://doi.org/10.1523/JNEUROSCI.3989-14.2015>
- Zilles, K., Bacha-Trams, M., Palomero-Gallagher, N., Amunts, K., & Friederici, A. D. (2015). Common molecular basis of the sentence comprehension network revealed by neurotransmitter receptor fingerprints. *Cortex*, 63, 79–89.
<https://doi.org/10.1016/j.cortex.2014.07.007>
- Zylberberg, J., & Strowbridge, B. W. (2017). Mechanisms of persistent activity in cortical circuits: Possible neural substrates for working memory. *Annual Review of Neuroscience*, 40, 603–627. <https://doi.org/10.1146/annurev-neuro-070815-014006>

Glossary

Action potential: brief electrical pulse (spike), with a generic shape and duration (1–2 ms) generated at the cell body when a threshold is exceeded. It propagates down the axon which connects to the dendrites of other neurons through one or more synapses.

Conductance: the ease with which an electric current flows through an object or material; the inverse of resistance.

Dendrites: branched, tree-like structure protruding from the cell body of a neuron that integrates synaptic input to change the neuron's membrane potential.

Dynamical variable: a physical quantity whose numerical value changes over time, describing some aspect of the system's state, e.g., the membrane potential of a neuron, or the conductance of a particular synapse.

Engram: a basic unit of information stored in long-term memory, e.g., a phoneme, word, or idiomatic expression.

G protein: protein that transmits signals from the exterior to the interior of a cell, acting as a molecular switch.

Homeostasis: when physiological variables deviate from a pre-set range of values, self-correcting feedback restores a dynamic equilibrium to retain stability.

Laminar structure: neocortical organization into six layers with characteristic connectivity within and between layers that forms cortical microcircuits.

Monoamine: class of neurotransmitters that alter the processing characteristics of entire circuits beyond the single synapse, e.g., dopamine, serotonin, histamine.

Neuropeptide: signaling molecule, or chemical messenger, that diffuses over broad areas and modulates neural activity.

Neurotransmitter: molecule that transmits signals across a chemical synapse from one neuron to another, e.g., glutamate or γ -aminobutyric acid (GABA).

Point neuron: mathematical model of a biological neuron that lumps all neuronal structure into a single, homogeneous compartment that receives input signals and generate output spikes.

Poisson process: a stochastic event process where random events are independent of each other and the time between events follows an exponential distribution.

Receptor: transmembrane protein that is activated by a neurotransmitter and regulates the activity of synaptic ion channels across the cell membrane.

Rheobase: the minimal current amplitude of infinite duration that results in the discharge of an action potential.

Spike-timing dependent plasticity: adaptive mechanism that adjusts the strength of synapses based on the relative timing of a neuron's input and output spikes, leading to, e.g., long-term synaptic potentiation (LTP) or depression (LTD).

Synaptic motif: synaptic connectivity pattern involving a small number of neurons, e.g., the bidirectional coupling between excitatory cells, or feed-forward inhibition.

Timescale: a characteristic time span within which a particular change in a dynamical variable takes place.

Unsupervised learning: innate, self-organized form of learning that detects patterns in unlabeled input without explicit instruction.

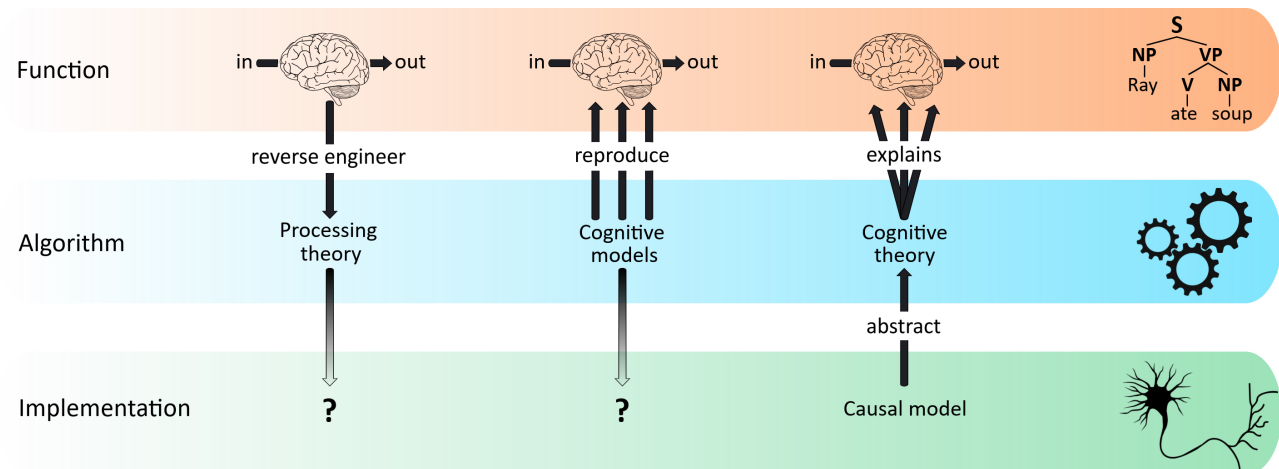


Figure 1

Three approaches towards understanding language as a cognitive system. Cognitive information processing systems can be described at different levels of explanation, here exemplified by the functional, algorithmic and implementational levels (Marr, 1982). A complete understanding of such a system would allow us to traverse seamlessly between levels in all directions. Although the three levels will have to be augmented with additional ones (Churchland & Sejnowski, 1988; Tanenbaum & Austin, 2013), this broad distinction has been fruitful in partitioning the problem space. This explanatory challenge can be approached in different ways. Experimental language science has attempted to infer processing theories from observed input-output relations (left). Cognitive modeling has proposed a large array of algorithms that can each reproduce some aspects of these relations (center). Causal modeling starts from neurobiological principles to synthesize an explanatory language model which is, first and foremost, a model of the system itself (right). Ideally, such a model will eventually explain all behavioral data generated by the system.

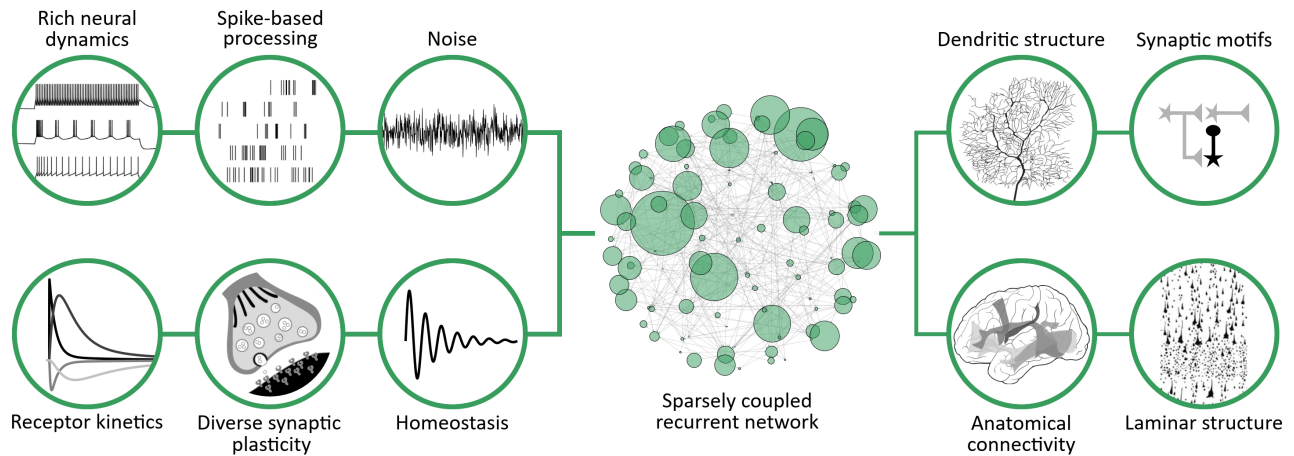


Figure 2

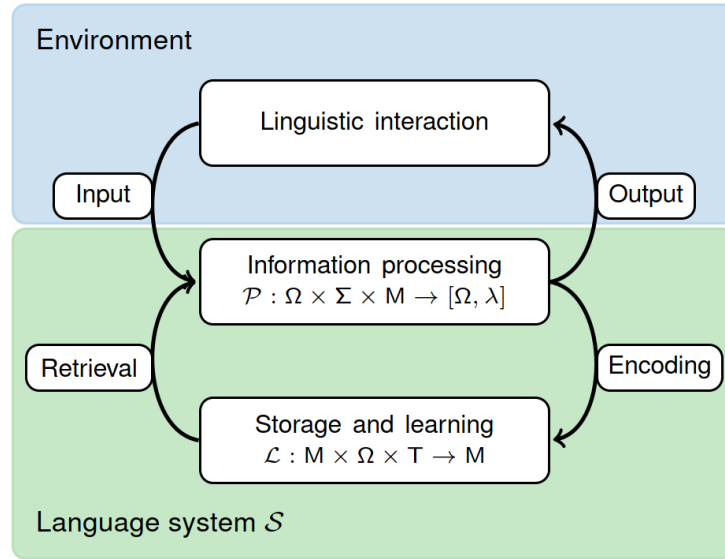
First principles of neurobiology. Features of the nervous system that are largely uncontroversial in neurobiology form the basis of causal language models. These can be viewed as boundary conditions that constrain proposed mechanisms for language processing. Causal modeling seeks to understand the computational role of these features in relation to language processing and integrate the implementational level with the algorithmic and functional levels of description.

Box 1. Causal modeling toolbox

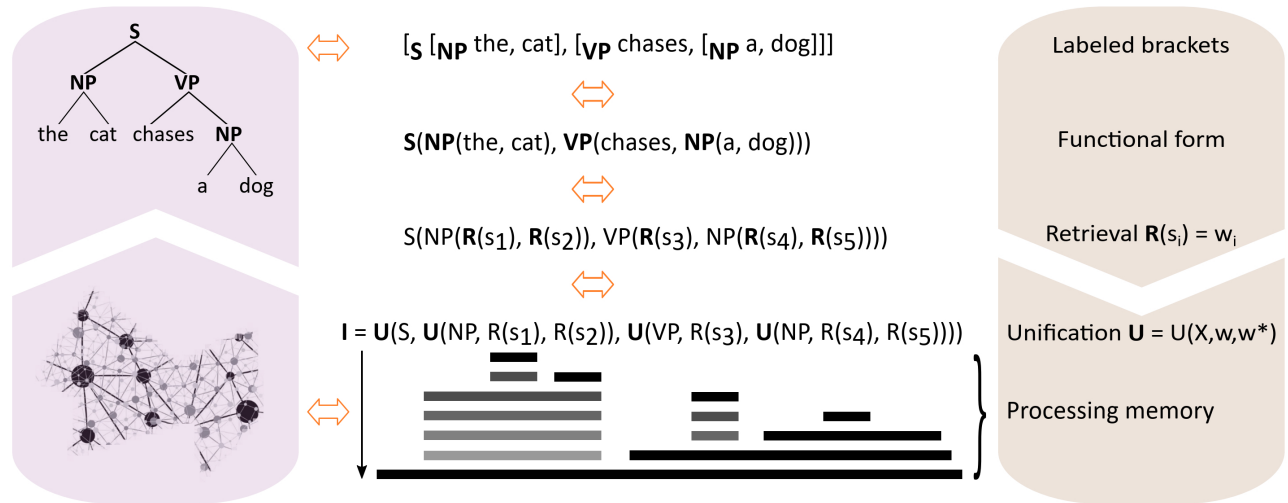
Although causal models operate at the implementational level, the aim is not to replicate reality in all its complexity. Instead, physiological processes are modeled in a phenomenologically effective manner. For many of the neurobiological features in Figure 2, *reduced mathematical models* exist from which causal networks of language function can be assembled, largely in the form of systems of coupled differential equations.

The distribution of cortical spikes can, under suitable circumstances, be approximated by **Poisson processes** (Softky & Koch, 1993) to *encode input* as frozen noise. This is an example of how one can create a spatio-temporal code for linguistic units which carries more information than a rate-based code (Duarte et al., 2018; Uhlmann, 2020). The two-dimensional, *adaptive-exponential neuron* is able to produce a wide range of firing patterns (Brette & Gerstner, 2005) and accurately predicts *in vitro* spike times (Rossant et al., 2011). *Synapses* can be modeled as alpha-functions or the difference between two exponentials that describe the rise and decay times of post-synaptic currents (Roth & van Rossum, 2009) and conductance-based coupling supports realistic population dynamics (Cavallari et al., 2014). Event-driven simulation can be used to efficiently model axonal delays for long-range connectivity patterns. *Short-term synaptic facilitation* and depression is modeled in terms of **neurotransmitter** release probability and depletion (Markram et al., 1998) and this mechanism has been implicated in working memory function (Mongillo et al., 2008). Excitatory long-term potentiation and depression are conceptualized as Hebbian **spike-timing dependent plasticity** (STDP). Several similar formalisms exist which are based, e.g., on triplets of spikes (Pfister & Gerstner, 2006), or on pre- and post-synaptic voltage traces (Clopath et al., 2010). The latter rule allows for strong bidirectional potentiation which has been observed experimentally. To counteract dynamic instability due to STDP, *inhibitory plasticity* acts on inhibitory synapses to maintain a target firing rate (Luz & Shamir, 2012; Vogels et al., 2011). This form of plasticity also establishes a local balance between excitatory and inhibitory synaptic inputs to each neuron and is conducive to achieving asynchronous, irregular spiking activity which plays an important role in cortical information processing (Herstel & Wieringa, 2021; van Vreeswijk & Sompolinsky, 1996). *Synaptic normalization* is another homeostatic principle which counteracts uncontrolled synaptic growth due to STDP while preserving synaptic specificity (Turrigiano, 2008). On longer timescales, relevance signaling and synaptic tagging models have been developed that prevent overwriting and enable memory consolidation (Clopath et al., 2008; Ding et al., 2022; Ziegler et al., 2015). What has been missing from this inventory of neurobiological components, until recently, are computationally efficient multi-compartmental neuron models, capable of reproducing non-linear dendritic integration effects that have been described experimentally (Koch, 1999; London & Häusser, 2005; Payeur et al., 2019; Poirazi & Papoutsis, 2020). The *Tripod neuron* proposes a structural reduction of the dendritic tree to fill this gap and can now be used to investigate the functional role of dendritic integration in large networks (Quaresima et al., 2022).

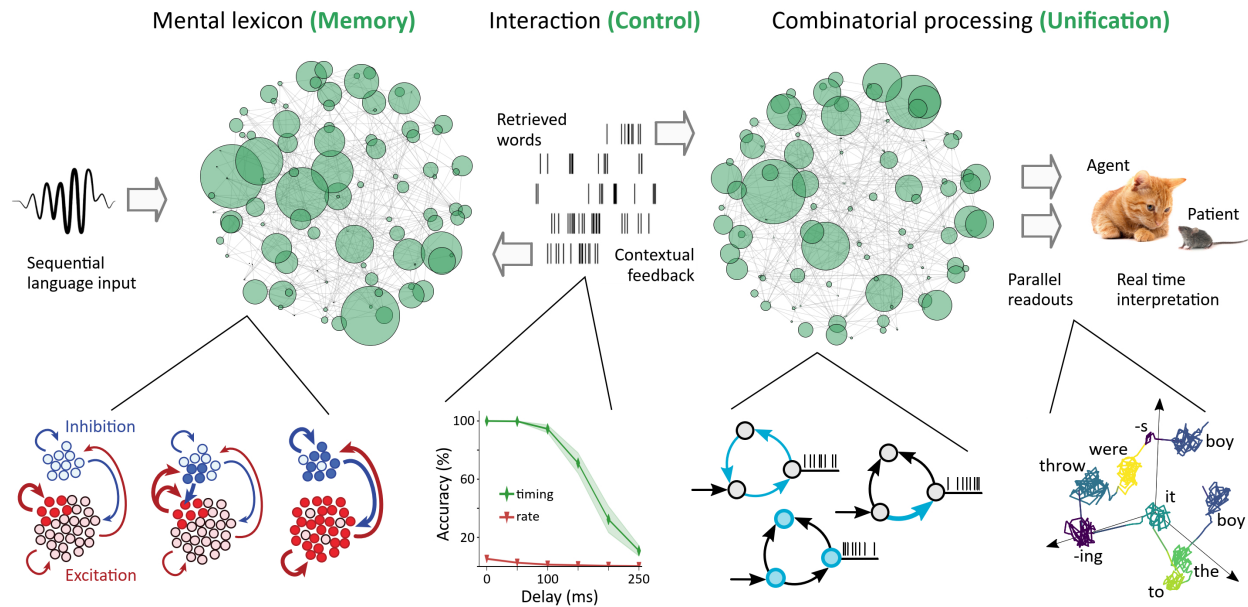
Causal language modeling is further supported by flexible, high-level spiking network simulators (Gewaltig & Diesmann, 2007; Stimberg et al., 2019), code-sharing platforms (McDougal et al., 2017) and programming languages for high performance scientific computing (Bezanson et al., 2017).

**Figure 3**

Schematic of an adaptive information processing system S for language. Based on input from the system embedding (environment; an element of Σ), the current state (an element of Ω), and current parameters in the model space \mathcal{M} , the processing dynamics \mathcal{P} traces out a trajectory in neural state space and returns language output λ . The learning mechanisms \mathcal{L} are coupled to \mathcal{P} creating continuous cycles of information encoding into, and retrieval from, memory that operates on multiple timescales for short-term and long-term storage as well as development. For instance, in the case of ontogenesis, \mathcal{L} implements developmental processes and genetically-guided maturation dependent on time \mathcal{T} , while \mathcal{P} instantiates the parsing capacity that evolves towards adult competence as a function of \mathcal{L} 's trajectory through the model space \mathcal{M} . On shorter timescales, \mathcal{L} implements an active processing memory and because the form of \mathcal{L} is structurally similar to \mathcal{P} , it is possible that learning and memory mechanisms are actively computing on relevant timescales as well (e.g., in transforming episodic memories into general world knowledge as a consequence of repetition during consolidation). Figure adapted from Petersson and Hagoort (2012).

**Figure 4**

Translating linguistic hierarchy into neural processing. Phrase structure trees are rewritten in labeled bracket notation where brackets correspond to nodes in the tree and labels indicate the category of nodes (orange arrows represent equivalence). Labeled brackets can be expressed functionally as $NP(\text{the, cat})$, and similarly for other phrasal categories in the example sentence. Words that enter into these function calls are retrieved from the mental lexicon by an operator R that incrementally maps speech sounds s_i onto word representations w_i . A parameterized function U (unification) is introduced that takes three arguments, a phrasal category and two partial interpretations w, w^* that have either been retrieved by R or computed by previous actions of U . To establish sentence meaning, nested function calls to U are executed in the correct order as soon as relevant information becomes available (immediacy principle) and the output of U corresponds to the interpretation I of an utterance. During this procedure, lexical items as well as partial interpretations previously computed by U have to be kept in processing memory until they are being integrated. Processing memory also keeps track of which components have already been unified, and when, in order to carry out potential revision. Grey-scale horizontal bars show the lifetime of information content temporarily held in memory at each processing step, the vertical arrow indicates logical time.

**Figure 5**

Core computational machinery for language processing. The cognitive architecture for language consists of a mental lexicon for the encoding, maintenance and retrieval of words and a unification network for combinatorial processing. Both components require memory on long and short timescales to different degrees and their interaction is a form of reciprocal control. Downstream readouts project the neural states of unification onto a semantic interpretation in real time. The distinction between memory, unification and control is purely functional, it is not a claim about anatomical localizability. Any computational system, whether neural or classical, implements these components in one way or another. Insets from left to right: word representations in the mental lexicon, or **engrams**, are strongly coupled cell assemblies, recruiting excitatory or inhibitory synapses, or both (bold arrows); figure adapted from Hennequin et al. (2017). Retrieved words are encoded as spike trains that drive unification and information content is better preserved in the timing of spikes than in spike rates, even in the presence of noise; figure adapted from Uhlmann (2020). Processing memory for unification may be implemented neurobiologically as network attractors, short-lived synaptic facilitation, or intra-cellular adaptation that transiently changes neuronal excitability (top to bottom). History-dependent processing in unification, where the current state is folded together with incoming input, separates multiple occurrences of the same word (here “boy”) in neural state space, and this can be used to establish binding relations between words and their semantic roles; figure adapted from Fitz et al. (2020).

Box 2. Methodological road map

Causal modeling initially puts priority on neurobiological realism over fit with behavioral data. Therefore, a first step is to create models from neurobiological components parts that can accomplish core computational tasks involved in language processing at the algorithmic level. Network features should comply with functionally relevant neurophysiological measurements. These include the electrophysiological properties of different neuron types (e.g., their resting potential, **rheobase**, membrane time constant, spike threshold, etc., that can be obtained from databases such as NeuroElectro (Tripathy et al., 2014) and the Allen Brain Map¹⁰), spontaneous and evoked firing rates (Attwell & Laughlin, 2001; Roxin et al., 2011), the quantized range of synaptic conductances (Bartol Jr et al., 2015), the ratio of excitation and inhibition (Abeles, 1991; Xue et al., 2014), and the distribution of major **receptor** types across regions in the language network (Duarte et al., 2017; Zilles et al., 2015). Language models with these characteristics have *face validity* since they are grounded in experimental neurobiology. This approach applies equally to networks of any spatial scale, including larger-scale neocortical or cortico-striatal networks (Haber, 2016; Mountcastle, 1997; Shepherd, 2004). In each case, the connectivity matrix would be structured into blocks with specified neuron types and local connectivity as well as specific between-region connectivity (cf. also Pulvermüller et al., 2021).

Key language tasks include, among others, the transduction of auditory signals onto equivalence classes (phonemes), the retrieval of lexical features (semantic, morphosyntactic, etc.) from these units of speech, and the integration of recognized words into a sentence-level interpretation (semantic dependency structure). To gauge task performance, simple parallel readout classifiers can be used as a *measurement device* that maps non-linear circuit activity onto linguistic categories (Buzsáki, 2010; Rigotti et al., 2013). Thus, readouts are a diagnostic tool to probe whether a given dynamical system can be harnessed to compute linguistic functions. The neurobiological features of this system can then be manipulated (another meaning of *causal modeling*) and their computational contribution can be determined through model comparisons as a method of investigation (Duarte & Morrison, 2019; Fitz et al., 2020; Uhlmann, 2020). Importantly, failure to achieve these language tasks is *inherently meaningful* because it points directly to missing neurobiological features that might be important for language processing. In addition, our current best models of neurobiological components may have to be refined or extended in light of new empirical evidence while the causal modeling framework does not need to be questioned as such.

Once a basic neurobiological language model has been established, causal modeling can begin to bridge into empirical data and linguistic behavior. For instance, local field potentials can be synthesized from peri-synaptic activity in simulated spiking networks (Hagen et al., 2016; Mazzoni et al., 2015) to connect causal models to ECoG, EEG and MEG data. In similar vein, hemodynamic response models have been proposed to link *in silico* network activity to fMRI data (Bonaiuto & Arbib, 2014). These methods can be used to relate causal models to functional neuroimaging. This endeavour also involves statistical approaches to quantifying single-neuron and population dynamics (Kass et al., 2018; Saxena & Cunningham, 2019) and the representational analysis of biological networks (Barrett et al., 2019). Novel techniques for analytic synthesis need to be developed that allow the abstraction of adaptive dynamical systems to discretized combinatorial models.

Causal modeling advances from neurobiological models of algorithmic capacities to neuroimaging data and linguistic behavior. Through incremental model refinement, the core objective is to uncover the computational role of neurobiological features and synthesize a computational neurobiology of language across levels of explanation.

Box 3. Open questions

- What are the elementary units of language in neurobiological terms (e.g., phonemes, syllables, words, phrases, clauses, semantic roles, event structure)? Which neural data structures encode these units and their composition, and how can these data structures be identified through causal modeling?
- What is the functional role of brain structure in language processing across spatial scales, including structure in the dendritic tree of neurons, laminar structure in cortical microcircuits, and connectivity structure between brain regions in the perisylvian language network?
- What is the neurobiological correlate of processing memory for unification? How does this system support temporal integration, the resolution of non-adjacent dependencies, and recursive function calls for compositional processing? How are intermediate processing outcomes stored, retrieved at the right point in time, and broadcast to where they will be used next?
- How is prior knowledge of language expressed within the neurobiological infrastructure of the language-ready brain and what is unique about human neurobiology that enables language in the first place? Causal modeling is ideally suited to test specific hypotheses concerning, e.g., dendritic morphology, cytoarchitectonic composition, receptor-architectonic fingerprints, and anatomical connectivity.
- What is the structure of words stored in the mental lexicon and how does it enable combinatorial sentence-level processing in biological networks? What kinds of representations are supported by the underlying neurobiology? How are they encoded and maintained in long-term memory in the presence of noise and ongoing plasticity, and how is the feature structure of words computed from partial cues?
- How is a language-specific mental lexicon acquired given the weak, local neurophysiological learning mechanisms currently known, and how does learning interact with innate structure during acquisition?
- The complexity of neurophysiology demands reduced mathematical models that abstract away from, e.g., ion channels and the molecular machinery of synapses. What is the appropriate level of reduction that is computationally feasible while still being informative at the algorithmic level?