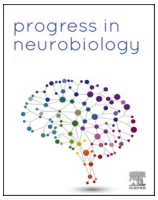
Progress in Neurobiology 230 (2023) 102511

Contents lists available at ScienceDirect

Progress in Neurobiology

journal homepage: www.elsevier.com/locate/pneurobio



Neurobiological mechanisms for language, symbols and concepts: Clues from brain-constrained deep neural networks

Friedemann Pulvermüller [a,b,c,d,*]

[a] Brain Language Laboratory, Department of Philosophy and Humanities, WE4, Freie Universität Berlin, 14195 Berlin, Germany
[b] Berlin School of Mind and Brain, Humboldt Universität zu Berlin, 10099 Berlin, Germany
[c] Einstein Center for Neurosciences Berlin, 10117 Berlin, Germany
[d] Cluster of Excellence 'Matters of Activity', Humboldt Universität zu Berlin, 10099 Berlin, Germany

ARTICLE INFO

Keywords:
Neurocognition
Neurocomputation
Semantics
Language learning
Deep neural network
Brain-constrained model

ABSTRACT

Neural networks are successfully used to imitate and model cognitive processes. However, to provide clues about the neurobiological mechanisms enabling human cognition, these models need to mimic the structure and function of real brains. Brain-constrained networks differ from classic neural networks by implementing brain similarities at different scales, ranging from the micro- and mesoscopic levels of neuronal function, local neuronal links and circuit interaction to large-scale anatomical structure and between-area connectivity. This review shows how brain-constrained neural networks can be applied to study *in silico* the formation of mechanisms for symbol and concept processing and to work towards neurobiological explanations of specifically human cognitive abilities. These include verbal working memory and learning of large vocabularies of symbols, semantic binding carried by specific areas of cortex, attention focusing and modulation driven by symbol type, and the acquisition of concrete and abstract concepts partly influenced by symbols. Neuronal assembly activity in the networks is analyzed to deliver putative mechanistic correlates of higher cognitive processes and to develop candidate explanations founded in established neurobiological principles.

1. Introduction

The brain mechanisms of the most advanced cognitive abilities, including language, symbol processing and conceptual thinking, have successfully been investigated throughout the last three decades. We now have substantial knowledge about which brain areas are necessary for different aspects of language processing and which regions 'light up' when meaningful symbols or concepts are being understood or produced. However, there is still a lack of understanding of the precise mechanisms that implement or realize language and conceptual thought at the level of neurons and neuronal assemblies. The mechanistic key question to address is how large populations of nerve cells interact in enabling human language use and meaningful communicative interaction.

Unfortunately, most current brain language models do not answer this question, but rather remain at the level of functional descriptions of (by assumption partly independent) processing components, which are assigned to brain structures, for example to specific cortical areas or fiber bundles interconnecting cortical areas. Such *co-labeling* of linguistic processing components and brain parts does not address the question how sets of neural units interact with each other when language processing takes place and, crucially, cannot address the questions of *why* and *how* language and symbolic functions come about or *why* they are bound to specific cerebral locations and activity dynamics.

One novel way to approach the question of how neuronal populations interact in symbolic, conceptual and linguistic processing is to try it out using neural models containing artificial devices similar to nerve cells. This strategy follows the idea that understanding a mechanism requires the ability to engineer it – which is concisely expressed by Feynman's famous remark "What I cannot create, I do not understand". Neural network models of language and conceptual processing have been extremely successful in simulating performance on specific linguistic and cognitive tasks, for example classification of object pictures into semantic categories, such as those designated by the words "face" or "animal", syntactic parsing, or translation from one language into another; convolutional and recurrent deep neural networks have even reached human-like performance on several such tasks (Kietzmann et al., 2019a; LeCun et al., 2015; Linzen and Baroni, 2021). More

* Corresponding author at: Brain Language Laboratory, Department of Philosophy and Humanities, WE4, Freie Universität Berlin, 14195 Berlin, Germany.
E-mail address: friedemann.pulvermuller@fu-berlin.de.

https://doi.org/10.1016/j.pneurobio.2023.102511
Received 3 December 2022; Received in revised form 2 May 2023; Accepted 18 July 2023
Available online 22 July 2023

recently, deep neural networks, some of them optimized by incorporating new fast and efficient algorithms, also allowed for model training and testing with huge data sets. For example, variational autoencoders (Higgins et al., 2016) and generative pretrained transformers (Vaswani et al., 2017) not only succeeded in modelling image analysis and language processing and generation at human performance levels but even successfully modeled data recorded from single neurons or whole brain neuroimaging (Caucheteux et al., 2023; Schrimpf et al., 2021).

However, it is one thing to build an algorithm or machine that approximates human cognitive performance, but a different one to find out in which way the human brain solves the related tasks (for discussion of this point, see part 7.1 below). Most 'neural' models are difficult to interpret in terms of brain mechanisms, because they lack commonalities and similarities to brain structure and function. For example, these networks may consist of a linear lineup of 'layers' or 'areas', which does not capture the complex connectivity structure between the areas of the human cortex. Likewise, connections between the artificial neurons within each 'layer' are not constrained by local connections in cortex, and the learning mechanisms applied are also sometimes different from what can be inferred from synaptic dynamics during biological learning. Because of these dissimilarities, it is difficult to interpret the achievements of neural network research in terms of brain mechanisms. Due to this shortcoming, researchers have suggested to constrain neural network modelling by multi-level information about the structure and function of the brain, so as to assure, or, to put it more moderately, to make it more likely, that network-immanent mechanisms resemble those in the real neuronal tissue (Deco et al., 2013; Dwivedi et al., 2021; Hahn et al., 2019; Kumar et al., 2010; O'Reilly, 1998; Palm, 2016; Pulvermüller et al., 2014; Pulvermüller et al., 2021; van Albada et al., 2022; Wennekers et al., 2006).

1.1. Aims and scope of this review

There are many eminent and exciting, hitherto unanswered questions about specifically human higher cognitive abilities, including those to use and understand language, symbols, meanings and concepts, that need addressing in neuromechanistic terms: How can humans build huge vocabularies of tens of thousands of symbols that far exceed those of their closest relatives, non-human primates, who typically use fifty to one hundred? Based on which mechanisms do infants learn the relationship between symbols and the referent objects and actions these symbols are used to communicate about? And how can infants learn this link with surprising speed, after only a few learning events? These questions will be addressed in Section 4, after explaining the brain-constrained modelling approach (Section 2) and discussing the nature of cognitive 'representations' emerging in neural networks (Section 3). Even more fundamental cognitive questions will be addressed in Sections 5 and 6: To what degree is language beneficial and even essential for building concepts, and what are the neuronal mechanisms underlying any such causal link? And how can symbols and language drive other cognitive functions, including attention to objects and memory? This requires a specification of the nature of the meaning-related or semantic links between information about symbols and 'the world', which differ between symbol types, in particular between so-called 'proper names' specific to one object and more general terms related to whole categories of similar entities (Section 5). Two other symbol and concept types are in the focus of Section 6, concrete and abstract ones. What is the mechanistic difference between concrete and abstract concepts, how might the learning of symbols for these concepts influence their representations, and how can any such differences be explained?

In sum, this paper will first highlight and discuss a set of constraints that can be applied to make sure that model networks show specific similarities to the real (human) brain (Section 2). It will then be asked which type of neural mechanism or 'representation' emerges in brain-constrained networks during the learning of symbols, thereby addressing the discrete or distributed nature of their neural correlates (Section 3). To illustrate brain-constrained models and the explanations they provide, Section 4 presents three case studies covering, respectively, the structural and functional correlates of verbal working memory and the vocabulary build-up it enables (4.1), the binding of information about a symbol and its typical referent (4.2) and the surprising speed with which form and meaning are mapped in early language acquisition (4.3). Section 5 addresses the mechanisms of form-meaning binding and its plasticity when different kinds of referential expressions, proper names and category terms, are learnt. This part will also address the sophisticated mechanisms by which language influences attention and flexibly directs it towards specific features of objects. In Section 6, the putative biological basis of the formation of concrete and abstract concepts will be in focus, along with the mechanisms underlying concrete and abstract semantic learning in context of symbols. Questions about the linguistic influences on concept formation and processing will be covered throughout. The last part, Section 7, will discuss perspectives and limitations of brain-constrained cognitive modelling.

When addressing these issues one by one, mechanistic models of the observed phenomena will be presented, which are rooted in neurocomputational modelling work with brain-constrained neural networks. These models will be used to derive biologically founded mechanistic explanations of the issues addressed. The aim is to show, using a broad range of questions about concepts, symbols and their meaning, how a neurobiological foundation and precise neurocomputational implementation can help making cognitive science and neurobiology an explanatory field of investigation.

2. Biological constraints, model validation and neurocognitive explanation

As mentioned, the need to make neural network models more neurobiologically realistic is well established (Deco et al., 2011, 2013; Dwivedi et al., 2021; Hahn et al., 2019; Henningsen-Schomers et al., 2023; Kumar et al., 2010; O'Reilly, 1998; Palm, 1982, 2016; Pulvermüller et al., 2014; van Albada et al., 2022; Wennekers et al., 2006). But which features of neural networks can be relevant for judging their similarity to brain structure and function? Below is a brief description of 7 important constraints, which are discussed in greater depth in a recent publication (Pulvermüller et al., 2021). Firstly, neurobiological networks are composed of *nerve cells* and a model can mimic the functionality of these elementary neuronal processing units. This constraint is met by most neural networks, as the processing units, which these networks consist of, are functional units similar to real nerve cells, which receive inputs from other such units, integrate them and compute an output. Still, similarity between real and artificial 'neurons' can be smaller or greater (see below, Rojas, 2013; Gerstner and Naud, 2009). Secondly, one of the key features of neuronal function is *neuroplasticity*, the change of responsiveness of a neuron due to processing in the past. Some learning mechanisms are only loosely related to possible biologically implemented learning algorithms (e.g., error-backpropagation), whereas others faithfully follow the dynamics of synaptic connections between neurons (e.g., Hebbian learning, Lillicrap et al., 2020; O'Reilly, 1998). Therefore, at the microscopic neuronal level, the neuron model and the plasticity rules are important aspects and can be constrained to be as biologically realistic as possible or adequate. Thirdly, the mesoscopic level addresses the composition of *local neuronal circuits* of cells of different types, including both excitatory and inhibitory neurons, and the way they are connected with each other. Note that many neuron networks are subdivided into local parts – so-called 'layers' or 'areas' – in which no between-neuron connections exist, whereas others implement these local links at variable levels of biological realism (see, for example, Elman et al., 1996; Palm, 2016; van Albada et al., 2020). Local connectivity and interaction are closely related to the fourth constraint addressing activation dynamics, the way a set of neurons activates and deactivates. For example, after input to excitatory neurons, there will be an activation phase followed by an inhibition, due to the secondary

activation of inhibitory cells, and possibly ongoing oscillations. The interplay between excitatory and inhibitory dynamics can result in *regulation and control mechanisms* preventing the presence of excessively strong or weak activation levels, thus keeping the system within the bounds of functionality (Braitenberg, 1978; Deco and Rolls, 2005). Regulatory and control processes also exist above the mesoscopic level, at the level of interaction between larger brain structures such as cortical areas and subcortical nuclei, as, for example, between neocortex and the hippocampal formation or between cortex, basal ganglia and thalamus, which can be translated into networks at different levels of biological detail (Bibbig et al., 1995; Braitenberg, 1978; Dominey and Arbib, 1992; Fuller et al., 2019; Yuille and Geiger, 2003). At the macro-level, there is, fifth, the structuring of the brain in large brain parts and their subdivision into *sub-parts with different anatomical structure*, for example the areas of cortex, and, importantly and sixth, the *connectivity* between these. These macro-level constraints, area-subdivision and connectivity-structure implementation, are also realized to different degrees in different types of neural models, and are incorporated meticulously into so-called 'whole-brain models' (Deco et al., 2013; Deco et al., 2015; Hagmann et al., 2008; Vohryzek et al., 2023). In summary, there are six aspects and thus domains for implementing constraints: the neuron model, neuroplasticity and learning, local circuit composition and connectivity, regulation and control processes, and area structure and long-distance connectivity (for detailed discussion, see Pulvermüller et al., 2021). In addition, as these constraints are applied at different scales, their multi-level nature can be counted as an additional, seventh dimension of constraint. The constraints are illustrated graphically in Fig. 1.

It is important to note that, by saying that constraints can, and should, apply at these different levels, the 'tightness' or level of constraint is not yet specified. The neuron model constraint can, for example, be realized by using artificial neurons with graded responses, with more realistic integrate-and-fire dynamics, or even with implementation of neurochemical and biophysical detail including synaptic transmitter release and ion-channel dynamics. The resemblance to real neurons can therefore be close or more distant. Likewise, local and global circuits can be modeled with different numbers of neurons and areas, cell types and level of detail of their connectivity. Although one may wish the highest level of constraint to be applied for all dimensions, such maximally-constrained networks would be impossible to realize due to technical and resource-related limitations, including computing resources and time. For example, modelling a network of billions of neurons approximating the size of the human cortex with each neuron being realized by a detailed biophysical model would lead to excessive computation times on currently available devices. It is therefore necessary to choose levels of constraint thoughtfully by weighing the importance of different resource demanding constraints against each other.

When considering specific neural networks applied today, it is clear that most realize some of the constraints mentioned above. However, all seven constraints are implemented only rarely in one and the same model, although incorporating an additional one of these constraints increases the level of biological plausibility of a model. For example, most neural networks use sets of excitatory neurons unidirectionally connected between a sequence of subsets of their neurons, their 'layers' or 'areas'. This is rather far from the complex connectivity structure between the areas of cortex. Furthermore, there are frequently no inhibitory cells and no connections between the neurons within one 'area', so-called auto-associative links, and the connections between 'areas', the hetero-associative links, are all-to-all. These features contrast

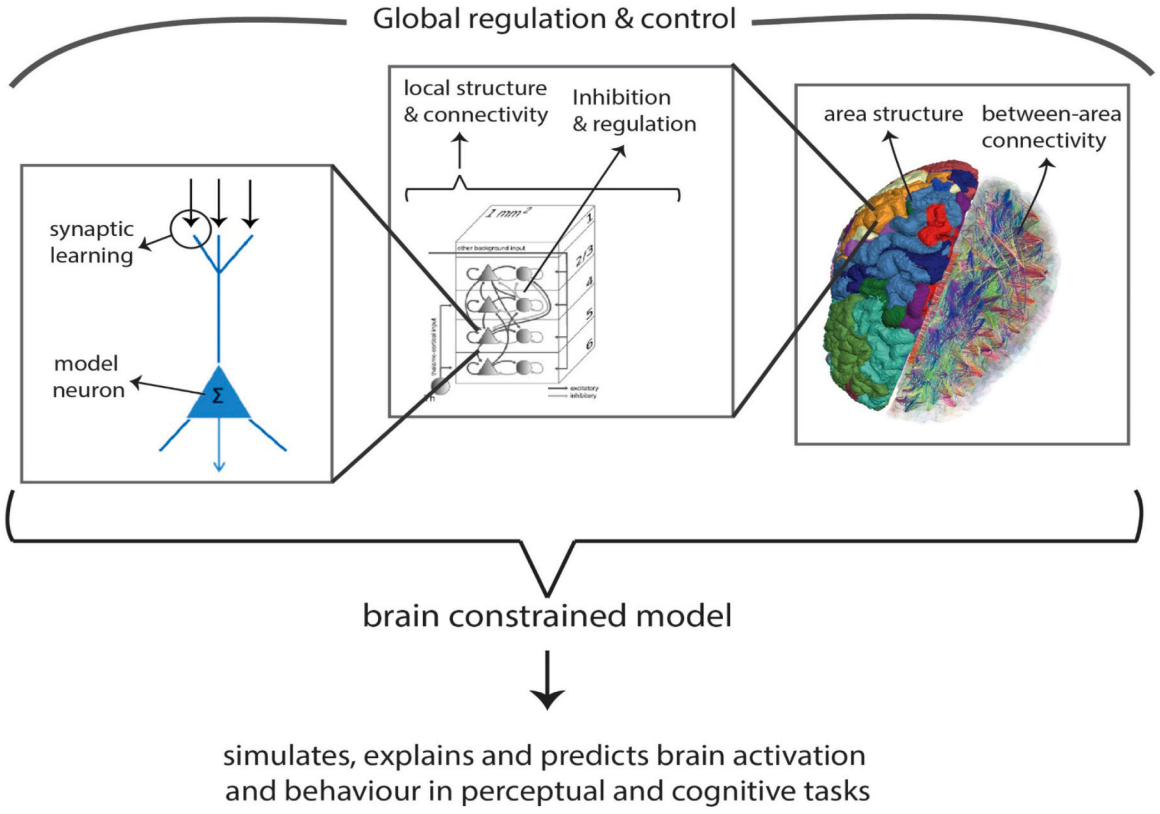


Fig. 1. Schematic illustration of constraints for making neural networks more similar to real brains. These include the micro-level of neuronal function and dynamics and the synaptic rules driving the plasticity of their connections, the mesoscopic level of connections and interactions between excitatory and inhibitory neurons within a local neuronal cluster and across such clusters within an area, and the macroscopic level of area structure and between-area connectivity. The models can be used as tools to predict and/or explain behavior and brain activity patterns. (The brain area and connectivity diagram on the right has been kindly provided by Rosario Tomasello and the local circuit diagram in the middle is reprinted, with permission, from Schmidt et al., 2018).

with the cortex's ample supply with inhibitory neurons and auto-associative connections between neurons in the same area and the well-known fact that the connections between most cortical areas are reciprocal, sparse and topographic, thus preserving neighborhood relationships (see, for example, Braitenberg and Schüz, 1998). Convolutional deep neural networks implement sparseness and topography of connections between adjacent layers/areas, and recurrent networks implement auto-associative connections. The implementation of sparse connectivity, topographical projections and recurrence can be seen as movements towards neurobiological reality (see, for example, Kietzmann et al., 2019b; LeCun et al., 2015), although other dissimilarities may still persist. As mentioned, between-area connections are not realistically implemented in most layered model networks. In contrast, whole brain networks implement the between-area connectivity structure revealed by experimental studies. However, in this case, the typical approach is to model one area by one neuronal element (or a set of differential equations modelling activity of a small neuron pool), thus missing opportunities for incorporating micro- and meso-level constraints (Pulvermüller et al., 2021; van Albada et al., 2022).

Therefore, in an attempt to make neural networks more neurobiologically realistic, it seems straightforward to realize all of the different constraints in the same model. This strategy may integrate some of the advantages of different types of classic neural networks, for example the between-area constraint of 'whole brain' networks and the local connectivity and neuroplasticity constraints realized by others (see, for example, Pulvermüller et al., 2021). Table 1 briefly summarizes in which way one example of a brain-constrained neural network addresses all 7 constraints discussed (see also Fig. 2 for network illustration and part 4.1 for detailed discussion, Schomers et al., 2017). The neuron model chosen is just weakly constrained, i.e. artificial mean-field neurons are used. Biologically realistic Hebbian learning is implemented using a biologically founded algorithm, which includes strengthening of the synapse between 2 neurons (long-term potentiation) when both are co-activating (if both surpass a pre-defined threshold) and weakening (long-term depression) when one of the connected neurons is active and the other one inactive. Regulation is realized through inhibitory loops at two levels, local (per neuron) and global (per area). Within each model area, there are sparse and random connections between excitatory neurons whose probabilities decrease with distance, and each excitatory neuron projects to one inhibitory cell which, in turn, projects back to it and its neighbors, so as to capture some aspects of local cortical connectivity (Braitenberg and Schüz, 1998; van Albada et al., 2022). The selection of cortical areas was guided by clinical observations demonstrating the importance of frontotemporal regions for language (including the perisylvian areas highlighted in shades of red and blue in Fig. 2, top left, Bates et al., 2003; Ivanova et al., 2018), and by connectivity studies suggesting a subdivision of these areas into primary, secondary sensory/motor and higher-order connector hub areas, the latter being characterized by comparatively many links to other areas (high connectivity 'degree', Sepulcre et al., 2012; van den Heuvel and Sporns, 2013). Individual between-area connections were motivated by connectivity and tractography studies (see Petrides and Pandya, 2009; Rilling, 2014). This model thereby addresses neurobiological constraints across levels, from neurons to local neuron clusters to areas and larger cortical regions.

Table 1
Illustration of the 7 constraints as realized in the model used by Schomers et al. (2017). For explanation, see text and Fig. 1.

constraint	realised in the model by
Multi level	Cortical columns, areas, perisylvian cortex (Fig. 2, top left)
Neuron model	Mean field neurons
Plasticity	Hebbian learning including synaptic strengthening (long-term potentiation) and weakening (long-term depression)
Regulation	Local and global regulation and control loops
Local connections	Sparse random neighborhood-biased excitatory links, inhibitory cells with local links (Fig. 2, bottom middle panel)
Area structure	6 areas in the left perisylvian cortex (Fig. 2, top left panel)
Global connectivity	Based on tracer and tractography studies in monkeys and tractography studies in humans (Fig. 2, left middle panel)

Such multi-level brain-constrained networks can be applied to address specific unanswered questions in the cognitive and brain sciences (see Introduction for examples). If the question addresses the nature of symbol representation in the human brain (Section 3) or the mechanisms underlying human-specific verbal working memory (4.1), the network can be used to simulate elementary word learning by the child (Fig. 2, right panels and caption). After learning, the strengthened neuron connections and newly-formed circuits can be extracted, documented and investigated, along with their activation dynamics and attractor states. Given sufficient 'internal' similarity of the network to the brain, these dynamics may provide clues about the neurobiological basis of the cognitive structures and processes targeted. To further scrutinize any tentative conclusions, constraints can be varied (for example by varying the neuron model or connectivity). Given the internal validity of the model provided by the constraints it observes, any process or representation developing in the model may appear as at least a candidate mechanism for neurocognitive theorizing. In addition, it is important that brain-constrained models are also validated externally, based on experimental data. For example, brain-constrained models similar to that of Fig. 2 and Table 1 have been validated using non-invasive neurophysiological recordings (EEG, MEG, Garagnani and Pulvermüller, 2011; Garagnani et al., 2008; Tomasello et al., 2017), neurometabolic activation (Garagnani and Pulvermüller, 2016; Tomasello et al., 2019), intracranial recordings (Pulvermüller and Garagnani, 2014), patterns of neuropsychological deficits due to brain disease (part 4.2) or learning patterns during infancy (Section 5).

Most importantly, however, these constrained networks can be used for developing novel neurobiological explanations of cognitive mechanisms, thus providing explanatory answers to questions about why specific processes are 'housed in different areas' or why discrete or distributed representations develop (cf. list of questions in part 1.1). Clearly, the model in itself does not provide an explanation, but the researcher can use it to develop one, or at least a candidate explanation. In doing so, the phenomena observed in the network need to be related to established biological principles, several of which are also manifest in the constraints applied. For determining the relevance of a given constraint for an explanation, it may be useful to vary its level and observe the effect (or lack thereof) on network function. In later sections of this review, such candidate explanations will be proposed for the emergence of human verbal working memory, for semantic learning of symbol-referent relationships, for the way language drives attention to objects and their features and for the formation of abstract concepts.

A sceptic may still object that, within a huge and highly complex artificial network, it might be hopeless to find clues about the mechanisms generating the complex neural interactions and dynamics putatively underlying the processing of concepts and thoughts – almost as hopeless as recording from and looking at the >10 billion neurons in cortex. Still, taking a more optimistic perspective, it may be possible to map and document emerging structure and functionality within a complex artificial network and relate these to both cognitive phenomena and underlying biological principles. The full accessibility of data from all model neurons may be beneficial in this endeavor. Sections 3 to 6 will now address recent brain theory and neurocomputational research aiming at such neurocognitive explanation.

3. Cognitive representations in mind and model

A range of exciting research projects focus on building models of cognitive mechanisms with more or less explicit reference to brain mechanisms. These address a broad range of cognitive domains, including memory, attention, object perception and categorization, logical reasoning and language along with conceptual and semantic

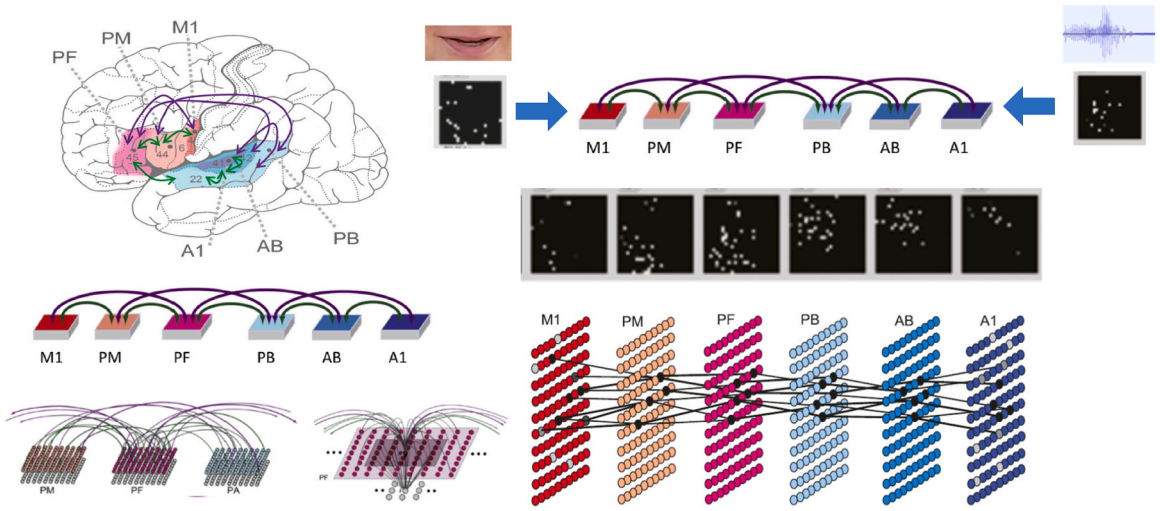


Fig. 2. Brain-constrained model of 6 areas of the left perisylvian cortex known to be of special relevance for language and symbol processing. The diagrams on the left show the network's structure and those on the right illustrate the simulation of the learning and processing of one word form. Top left: 6 areas are modeled, the inferior primary articulatory motor, M1, inferior premotor cortex, PM, and the inferior prefrontal area, PF, in the frontal lobe (in reddish colors) and primary auditory, A1, auditory belt, AB and parabelt, PB, in superior temporal cortex (in shades of blue). Middle left: Connectivity between the 6 model areas as suggested by tractography results. Bottom left: Illustration of the sparse, topographic and partly random connectivity between next and second-next neighbor areas. Bottom middle: Illustration of the sparse, topographic and partly random local connectivity within an area, between a given excitatory neuron (in the center) and other excitatory neurons (grey area) and between the excitatory cell, its inhibitory unit (shown at the bottom) and a smaller local neighborhood. Top right: To simulate the learning of a word form, a neuronal pattern representing articulatory features was activated in M1 and a pattern indexing acoustic features of the speech signal in A1. Activation was allowed to spread through the network and non-supervised Hebbian synaptic plasticity was applied. Middle right: After multiple such learning events, a strongly connected neuronal assembly has formed, which comprises neurons in all 6 areas, which activates as a whole upon later stimulation. Black squares show the six areas (from left to right, as in the left middle panel) and white dots the active neurons. Bottom right: Schematic illustration of a neuronal assembly spread across all 6 areas of the network. Small ovals represent neuronal units and lines strengthened connections between cell assembly neurons. For clarity, second-next between-area connections and within-area connections are omitted in this latter diagram.
(Panels on the left and the bottom right one are reprinted with permission from Garagnani et al., 2008; Pulvermüller and Garagnani, 2014; Schomers et al., 2017).

processing (e.g., Cazin et al., 2019; Constant et al., 2023; Dell, 1986; Dell et al., 1997; Dijkstra et al., 2019; Dominey and Inui, 2009; Drude et al., 2018; Eliasmith et al., 2012; Elman, 2005; Grainger and Jacobs, 1996; Henningsen-Schomers et al., 2023; Henningsen-Schomers and Pulvermüller, 2022; Huyck and Passmore, 2013; Jackson et al., 2021; Kietzmann et al., 2019a; Kriegeskorte and Diedrichsen, 2019; Lindsay et al., 2017; MacKay, 1987; Papadimitriou et al., 2020; Stefaniak et al., 2019; Tomasello et al., 2019; Verduzco-Flores et al., 2009). The following paragraphs will focus on work targeting the learning of symbols and concepts, with a focus on the question what kind of 'representations' develop in artificial neural networks and the brain.

3.1. *Semantic models: discrete or distributed?*

Building a mechanistic model of symbol processing and language was typically based on a priori decisions about the nature of these mechanisms. The most established theories in cognitive science assume that symbols, concepts and meanings are discretely represented in the mind and brain, so that there would be, for example, one specific representation for each word form linked to separate representations of its meaning, syntactic function, phonological structure and so on (see, for example, Levelt, 1989; Morton, 1969). Researchers in this tradition proposed neural implementations of this type of model, in which a given representation is realized as a locally-separate and discrete processing unit, which can be active or inactive at any point in time. These so-called 'localist' neural models implement discrete representations by neural units or 'nodes' for a given symbol form and further associated nodes for the symbol's phonological and semantic feature representations (see, for example, Dell, 1986; MacKay, 1987; McClelland and Elman, 1986). Representations of perceptions and motor acts are realized by additional units separate from symbolic representations.

The existence of discrete symbolic representations separate from perception and action processes has been called into question by philosophers, linguists and cognitive scientists (Engel et al., 2013; Gibson, 1979; Lakoff, 1987; Langacker, 1991; Varela et al., 1991). This position builds upon intrinsic links between abstract cognitive operations and their related perceptual antecedents as well as the actions by which these cognitive operations become manifest or expressed. Some have even argued that the existence of such intrinsic links to action knowledge makes a separate level of cognitive representations, in the sense of "context-neutral description[*s*] of object features", obsolete (for discussion, see Engel et al., 2013).

Converging with this 'anti-representational' (better: anti-discreteness) position, computational scientists argued that distributed network models, for example three-layer feedforward networks with or without additional memory mechanisms and multi-layer 'deep' neural networks, do not develop neural analogs to 'representations', in the sense of discrete symbol-related processing units, but give rise to dynamic distributed activity patterns instead (Elman, 2004; Elman et al., 1996; Farah and McClelland, 1991; McClelland and Rumelhart, 1985; Ralph et al., 2017; Rogers and McClelland, 2004; Westermann et al., 2006). Because, in this case, the same set of neuronal units may be involved in the processing of many different symbols, the correlates of symbols within these fully-distributed networks was proposed to be better captured by dynamic activation vectors across, for example, an entire network layer. Therefore, the lack of constant discrete representations within fully-distributed networks was seen as an advantage, which opens novel perspectives to capture semantic and functional similarities and differences between the same symbols used in different contexts, which may escape a discrete localist approach (Elman, 1990; Elman, 2004).

However, proponents of the discrete symbolic representational camp question this position, for example because it does not readily offer criteria for assessing whether a symbol used in slightly different contexts comes with the same or different meaning(s), and because it is unclear how abstract symbolic meaning can be built from, and separated from,

concrete action- and perception-related knowledge. In addition, a principal problem with fully-distributed dynamic vectors is patterns interference. When perceiving both a cat and a dog at the same time, neurons indexing aspects of both entities are active together and the resultant cumulative pattern of activity will be substantially different from both individual patterns, so that information about the individual objects gets lost. It therefore appears that the distributed dynamic activations have their advantages over localist representations (e.g., mapping of contextual semantic differences), but that the discreteness and separability of mechanisms for different concepts and symbols comes with other advantages instead (e.g., robustness against pattern interference. For a broader discussion of these and related issues, see Clahsen, 1999; Elman, 2005; Fodor and Pylyshyn, 1988; Marcus, 2018; Marcus, 2008; McClelland and Patterson, 2002; Pinker and Ullman, 2002).

3.2. Network correlates of symbolic form

After all, it is an empirical question whether, within a neuronal machinery, abstract cognition, concepts and symbols can be 'built from' action and perception and whether there are discrete processing units in the human brain that can be likened to word forms and their meaning. To find out, one may perform neurophysiological experiments aiming at defining the set of neurons that become activate when given symbols or words are being processed. Obviously, such experiments come with serious problems due to the vast number of vocabulary items (several 10.000s) and the monstrous number of neurons to be examined (>10 billion if only cortical neurons are considered). Therefore, defining the precise set of neurons in the human brain that is active specifically when processing a specific word form or concept appears as an undoable task. All we can provide is clues about specific neurons from a small part of the brain related to a small set of symbols (e.g., Creutzfeldt et al., 1989; Yi et al., 2019) or about large-scale activation patterns associated with sub-categories of symbols (e.g., Carota et al., 2017; Pulvermüller, 2013) or individual ones (Bouchard et al., 2013; Carota et al., 2021; Huth et al., 2016). These results indicate that action and perception knowledge may be activated in conceptual and semantic processing and may even be necessary for it (Barsalou, 2008; Binder and Desai, 2011; Kiefer and Pulvermüller, 2012; Pulvermüller, 2018b), although the question about necessity remains controversial (see, for example, Dreyer et al., 2020; Vannuscorps and Caramazza, 2016). The results do not offer strong implications for the existence of discrete processing units in the brain which could underlie and 'represent' symbols, word forms, concepts and meanings.

A different possibility to address the question about the existence or absence of discrete processing units is to perform brain-constrained network simulations and observe what kind of mechanism develops when symbols are learned by brain-like systems. One may argue that this strategy has already been followed, with the above-mentioned results. However, please recall that the fully-distributed networks typically applied to address this question used just a few of the constraints highlighted in Section 2. Could it be that the application of a broader range of biological constraints to the networks leads to the emergence of discrete processing units that show similarities to the representational units postulated by symbolic cognitive theories?

Garagnani and colleagues used a brain-constrained network mimicking six areas in frontal and temporal cortex close to the lateral or sylvian fissure – the perisylvian cortex – which is known to be particularly important for language (Fig. 2, upper left panel, Bates et al., 2003; Ivanova et al., 2018). This part of the brain will therefore be called the 'perisylvian language cortex', although it is clear that additional areas are important for language and that the left perisylvian areas carry processes different from language, too (Fedorenko and Thompson-Schill, 2014; Pulvermüller, 1999, 2018b; Tremblay and Dick, 2016). The frontal areas included articulatory primary and premotor cortex and adjacent inferior prefrontal sites and the temporal areas primary auditory, auditory belt and parabelt cortices. Bidirectional, sparse and topographic excitatory connections interlinked adjacent areas and an additional link between prefrontal and parabelt areas was added. Similarly sparse excitatory connections were established within each area along with local and area-specific inhibition mechanisms. Connection weights were subject to Hebbian unsupervised learning, whereby simultaneous activation yielded synaptic strengthening and either pre- or post-synaptic activity alone led to weight reduction (Artola and Singer, 1993; Tsumoto, 1992). The production of spoken syllables and spoken word forms was simulated by activating a selection of neurons in the model's 'articulatory motor' and 'auditory areas' (Fig. 2, top right diagram). The simulations were intended to mimic cortical processes that take place when the baby utters its first syllables and word-like utterances and perceives these self-produced elements acoustically. (This does in fact not lead to exactly simultaneous activation of articulatory and auditory neurons, but the network effect of slightly asynchronous activation onsets yields similar results.) The study revealed that the co-presentation of sparse random patterns of neuronal activity at the opposite ends of the network (articulatory and auditory areas) leads to activity spreading across the network and to the formation of specific circuits of strongly connected neurons for each of the probed syllables and word forms (Garagnani et al., 2008, 2009).

Note that such bi-directional activation spreading can only occur in networks imitating the known bi-directionality of most cortical between-area pathways, which contrasts with the unidirectional excitatory connections between the layers of most neural networks. Most importantly, the formation of a unique set of strongly connected neurons or cell assembly for each word form is not a trivial consequence of the neurocomputational learning regime. On the background of earlier simulations using parallel distributed 3-layer networks or deep networks with 6 or more layers, dynamic and fully-distributed activation patterns within each layer would have been expected, without showing discrete functionality (Elman, 2005). Other types of networks, for example so-called auto-encoders, yield single neural elements coding cognitively interesting information (Higgins et al., 2021). And the modular perspective long dominating the field of cognitive science suggests that speech production units and speech perception/comprehension mechanisms are built independently from each other, as separate and quasi-autonomous processing units. Against this background, the formation of neuronal circuits specific to words, which interlink articulatory and auditory information, appears as surprising. But which brain-like features of the network might be critical for the emergence of distributed cell assemblies? This issue will be addressed in part 3.3 below.

It is worthwhile to look more closely at the processes implemented in the network, along with the learning results the network model brought about. To simulate near-simultaneous production and perception of 'babbles' and early words, patterns of neural activity thought to determine specific articulatory movement sequences were 'injected' in the 'motor cortex area' at one end of the network (red area M1 in Fig. 2, top panel on the right). At the same time, the 'auditory cortex area' at the opposite end of the network (blue area A1) received a different neuronal activation pattern thought to code for specific features of the acoustic signals produced by the infant's articulation. The co-activation of 'articulatory motor' and 'auditory cortex' by specific patterns led to activation spreading forward and backward (from auditory to motor and back) throughout the network, also involving all 'higher' fronto-temporal areas connecting the sensory and motor fields. As a consequence, the connections between the neurons involved in this activation spreading strengthened their mutual connections, due to joint pre- and post-synaptic activation. The result was a specific set of strongly connected neurons for each articulatory-acoustic phonological pattern (an example is shown in Fig. 2, middle panel on the right). Notably, each of these neuronal assemblies was distributed across all 'areas' of the deep network. In addition to their distributedness, each cell assembly consisted of neuron members which were more strongly interlinked with

each other than to neurons outside the set, within the larger 6-area network. This strong internal connectivity made the cell assembly a functional unit. Therefore, each of the cell assemblies can be considered as the putative neuronal machinery that processes one specific word or symbol form.

As specific cell assemblies emerged for each word form, it appears that biologically-constrained networks tend to build processing units for individual word forms and symbols. Whether or not one wants to call these 'representations' is an open issue. Certainly, these processing units do not 'represent' in the sense that a person would have created them with the intention to stand for something else (as a physicist can 'represent' force by an 'F' in a formula); the possibility that readers might be misled by the term was a reason for (neuro-) philosophers to recommend dropping it (see, for example, Baker and Hacker, 1984; Bennett and Hacker, 2006). It was the stimulation patterns related to sensory and motor features of word forms that led to the development of neuronal assemblies in the network; insofar, there was no person or entity that caused the 'representation'. Neural aggregate formation simply happened, caused by activation of many neural elements and the neurobiological principle of Hebbian learning, which was effective in a neuroanatomically plausible architecture. However, as each neuronal assembly stands for a different word form or symbol, they can be considered to be similar to representations a programmer uses when writing code. In this sense, the circuit in the neural system represents symbols used 'in the world', by humans or artifacts. In order to avoid the potentially confusing aspects of the word "representation" in this context, it is possible to speak, instead, about the putative "machinery", "material basis" or "mechanism" of a word form or symbol. However, considering the above-mentioned data from brain-constrained modelling, these neuronal assemblies are *discrete* devices, insofar as (i) they can be activated as a whole and/or remain inactive, and (ii) the neuronal machineries of two different symbols are distinct sets of neuronal elements (see also Pulvermüller and Garagnani, 2014; Pulvermüller et al., 2014). Although these sets are distinct, they may overlap with each other and each set activation may differ from other activations, for example when a set of specific neurons are pre-activated by the preceding context. Therefore, the discrete and distributed nature of the cell assemblies bear the potential to model discrete symbol processes along with gradual context effects. It appears that this discrete-distributed nature may help to integrate advantages respectively claimed by the proponents of localist vs fully-distributed representations (Elman, 2005; Fodor and Pylyshyn, 1988; Marcus, 2018; Marcus, 2008; McClelland et al., 2010; Pinker, 1994; Plaut and Patterson, 2010). In this sense, the cell assembly mechanism provides 'representations' that are both discrete and distributed.

3.3. *Why would brain-constrained networks build discrete cognitive circuits?*

In the above described network simulation of symbol learning, activity in the periphery of the network model, in its 'sensory and motor areas', led to synaptic modification and neuroplastic learning across all 'areas' of a deep network architecture. This was surprising, because the Hebbian learning mechanism applied in these simulations is a local learning process only affecting the synaptic connection between two nerve cells, but not neurons further apart. To guarantee neural plasticity across layers, most previous modelling of cognitive learning with multi-layer networks used non-local learning rules, such as error-backpropagation or other 'gradient dependent' algorithms (Lillicrap et al., 2020; Richards et al., 2019; Rumelhart et al., 1986). These algorithms are called 'non-local' and 'supervised', because, after the to-be-learnt information is provided to the network and the network's performance is evaluated with reference to a desired output or 'teacher signal', the amount of 'error', that is, each neuron's contribution to any deviation from the desired output, is computed and fed back to all neuronal units. Subsequently, the weight of each synapse in the network is modified, depending on the amount of error it contributes. Therefore, this type of learning mechanism makes it possible to implement neuroplastic changes non-locally, at all neurons of a multi-layer network. Error-gradient dependent algorithms have been shown to be extremely efficient and they contributed to the great success of deep neural networks in modelling human-like performance (Dahl et al., 2012; Graves et al., 2013; Krizhevsky et al., 2012; Smit et al., 2021; Zhou et al., 2019). However, it has been suggested that this type of learning lacks biological plausibility (Lillicrap et al., 2020; O'Reilly, 1998) and the discussion of its possible biological correlates is still ongoing.

In the brain, Hebbian plasticity plays an important role for learning and knowledge accumulation and is indispensable for the acquisition of perceptual, cognitive and linguistic capacities (Kempter et al., 1999; Keysers and Gazzola, 2014; Palm, 1982; Pulvermüller, 1999; Rauschecker, 1991). Hebbian learning includes at least two mechanisms. First, neurons that fire together strengthen their mutual connection(s) and second, and equally importantly, neurons firing independently of each other weaken their links (cf. long-term potentiation and depression, Artola and Singer, 1993; Tsumoto, 1992). The timing of neuronal activations also comes into play, as the activation of the pre-synaptic neurons only after that of the post-synaptic one typically leads to synaptic weakening, too (cf. spike-timing dependent plasticity, Bi and Poo, 1998; Caporale and Dan, 2008; Gerstner et al., 1996). As Hebbian learning is a local process, it may appear as questionable whether it can bring about neuroplasticity across the areas of a multi-area network. For example, it is unclear how the connection between neurons in the middle of the network of Fig. 2 (e.g., from area PB and PF) can be influenced by information far away, in the periphery of the 'deep' network (area A1 and M1). And of course, the same question can be raised regarding connections between neurons located in multimodal cortical 'association' areas of the real cortex, which do not receive direct sensory input or provide motor output. Such considerations may have contributed to the preference of non-local supervised rules in simulations using multi-layer architectures.

However, as shown by the simulation study explained in part 3.2, Hebbian mechanisms – including the association and dissociation components – are in fact sufficient to account for learning across the many layers of a deep neural network with bi-directional connections (see, for example, Doursat and Bienenstock, 2007; Garagnani et al., 2007, 2008, 2009). The reason is the aforementioned spreading of activation. When a network with only weak links is stimulated, this first leads to reliable activation of the stimulated neurons exclusively. Nevertheless, with some likelihood, due to the probabilistic nature of neural activation and the omnipresence of noise in any brain-like network, the directly stimulated neurons will also activate some of the nerve cells they connect to, thus leading to neuronal co-activation and strengthening of some of the links (Doursat and Bienenstock, 2007). The chain of strongly linked neurons grows with further activations, so that, after some time, strongly connected local ensembles will have formed through which a wave of activity reliably spreads each time there is stimulation from outside (Doursat and Bienenstock, 2007). This chain formation takes place from both sides in the six-area network of Fig. 2 when neuron populations in the motor and auditory areas (M1, A1) are activated together, so that, after some learning, the two neuronal chains will meet in the central areas of the network and eventually join due to their co-activation (Garagnani et al., 2007, 2009). This is why the local rule has non-local effects. In the context of gradually emerging activation spreading, local Hebbian learning leads to formation of strongly connected neuronal sets spread out across a multi-area architecture.

The result of strongly interlinked processing units in distributed neuronal networks with brain-constrained structure and function contrasts with distributed 'representations' emerging in fully-distributed layered network models. As mentioned, Elman and others have strongly argued that neural networks do not include separable mechanisms for individual words, symbols, meanings, concepts and other linguistic and cognitive units. Instead, dynamic activation patterns

emerge which show now obvious relationship to symbolic engrams, as all of these patterns use the same set of neurons, including all 'nodes' of a given 'hidden layer', to code for representing and processing a given cognitive item. This does not rule out the possibility that specific neural elements in each 'layer' may contribute preferentially to the processing of one or a selection of engrams (see Wood, 1978; Wood, 1980), but such units appear to be rare and difficult to detect in fully-distributed networks, if their presence is not even just due to "subtle manipulation on the part of the experimenter designed to produce the desired effect" (Plaut, 1995). If so, what is the explanation for the absence of specific and discrete neural processing units for symbols, concepts and meanings in most layered neural networks and their presence in the above-mentioned biologically constrained networks?

One possibility is that Hebbian learning on its own provides the key. Garagnani and Pulvermüller indeed argued that the 'dissociation' term of Hebbian learning – the 'neurons out of sync delink' rule – accounts, in part, for the separation of neuronal assemblies which would otherwise become associated or at least strongly overlap (Garagnani et al., 2009). However, to what degree overlap reduction is driven by dissociation learning depends on the precise patterns of correlation and on the choice of learning parameters. Current network simulations indeed show some neuronal overlap of the processing units corresponding to different symbols. This observation suggests that, although the Hebbian principle may reduce cell assembly overlap, it cannot guarantee full separation.

Considering connectivity constraints, it is apparent that within-area excitatory and inhibitory connections are absent from most distributed layered cognitive networks, whereas these are implemented in brain-constrained networks. That co-activated neurons strengthen their mutual links not only across the 'areas' of the network but also within each area, appears to be another factor substantially contributing to the formation of strongly connected processing units for different symbols. Furthermore, the within-area inhibition provided by the area-immanent inhibitory neural units may contribute to the separation of different neuronal assemblies. Although each symbol-related neuronal set is distributed across the areas of the entire network, may vary in the precise way it activates when being stimulated and may overlap with several other such assemblies, the individual sets can easily be separated from neurons not significantly contributing to the processing of a given symbol (Garagnani et al., 2009). In essence, it appears likely that it is precisely the more biologically plausible inner architecture of the 'areas' of brain-constrained networks – in particular their excitatory and inhibitory within-area connections – which, conjoined with full Hebbian association and dissociation learning, gives rise to the emergence of the material correlates of specific symbolic forms.

It will be relevant to compare the effect of variable types of within- and between-area connectivity, along with that of different learning algorithms, on the formation of discrete circuits. If cell assembly formation persists under varying connectivity patterns, this would be evidence for a predominant role of Hebbian plasticity. If only specific types of local links (e.g., including local inhibitors) or between-area connections (e.g., reciprocal ones) yield discrete word-related circuits, the relevance of these structural features for an explanation of symbol representations would be evident. Variable learning rules may also influence the properties of emerging representation. These are but some explanation-related issues that await future research.

4. Brain-constrained modelling of symbolic and semantic mechanisms

The following paragraphs address basic steps in modelling symbols and semantic links using biologically constrained neural networks. Three 'case studies' of symbol processing in brain-constrained networks will be summarized. First, the specific connectivity structure between areas of the human perisylvian cortex, which is of great importance for language (Fig. 2, top left), will be used to work towards an explanation of verbal working memory and the learning of large vocabularies it enables. Second, the mapping of information about word forms and their referent objects and actions will be probed in biologically founded network models, thereby addressing the question why specific cortical areas are important for semantics generally, whereas other areas are relevant for semantic processing of particular semantic categories specifically. Third, it will be asked why the mapping of words and their referents can be very fast, being established within a few trials and even sometimes taking place already after one exposure. These examples are intended to propose basic biological mechanisms and explanations of symbol learning, symbol grounding and semantic mapping. These biologically founded mechanisms will lay the ground for the explanation of more sophisticated semantic phenomena addressed in subsequent sections.

4.1. Neuromechanistic basis of human-specific verbal working memory

One of the aforementioned eminent questions (see 1.1) addresses the species-specificity of mechanisms underlying symbolic systems and language. Only humans can acquire languages with huge vocabularies and flexible combination of atomic symbols into complex strings. Other primates lack this ability in spite of great resemblance of their brains to those of humans. A possible explanation of this fascinating human specificity comes from neurobiology and could be rooted in differences in cortical connectivity between primate brains (Ardesch et al., 2019; Barrett et al., 2020; Braunsdorf et al., 2021; Dick and Tremblay, 2012; Frey et al., 2014; Glasser and Rilling, 2008; Petrides and Pandya, 2009; Petrides et al., 2012; Rilling, 2014; Rilling et al., 2011; Rilling et al., 2008; Rilling and van den Heuvel, 2018; Thiebaut de Schotten et al., 2012).

A neurocomputational study used a brain-constrained model of the perisylvian areas of the left cortical hemisphere, the brain part most important for human language, and compared this model's performance to that of a 'monkey model' (Schomers et al., 2017). The study focus lay on a specific anatomical difference in connectivity structure between human brains and that of monkeys and apes. In contrast to quantitative differences, for example the increases of specific areas and the brain as a whole, connectivity structure shows a well-documented qualitative structural evolutionary change (Rilling, 2014). The dorsal connection between left inferior frontal and temporal cortex by way of a pathway called the arcuate fascicle is strongly developed only in humans, but much weaker in chimpanzees and quite weak in macaques. Over and above stronger vs weaker development of the arcuate fascicle, it appears that particular between-area connections are characteristic of the human brain. A largely linear line-up of areas holds for most primates, with connections from primary to secondary auditory areas, from there to multimodal superior-temporal cortex, and on to prefrontal, premotor and motor cortex. There are additional indirect links, which provide ample shortcuts to this next-neighbor structure, for example between premotor and superior-temporal and between auditory belt and prefrontal areas, and, interestingly, these are much more prominent in humans than in other primates (see, for example, Rilling et al., 2011; Rilling et al., 2008; Thiebaut de Schotten et al., 2012). These so-called 'jumping links' (Fig. 2, left middle panel) may provide more effective information exchange within the human left-perisylvian language cortex.

The simulation study by Schomers and colleagues showed that the increase in connectivity leads to more efficient learning of correlated patterns of articulatory and auditory information and, crucially, the formation of discrete distributed processing units that maintained their activation for some time after their stimulation. In contrast, sensorimotor patterns presented to the 'monkey model' lacking 'jumping links' produced neuronal assemblies which activated after stimulation but then lost activity quickly. The temporary bistability and activity maintenance of the cell assemblies of the human model can be interpreted as a biological correlate of verbal working memory, a prerequisite of building a large vocabulary of symbols and thus for learning a language

(Schomers et al., 2017).

These simulations using brain-constrained networks offer a first step towards an explicit neurobiological explanation of specifically human language skills. Note that this explanation is based on just one minimal difference in connectivity structure implemented as part of the between-area connectivity constraint. The networks, which were multiply constrained by biological features, developed verbal working memory only if their connectivity structure included the indirect between-area links characteristic of the human connectome. Just increasing connection strength in the 'monkey architecture' did not yield this effect. With 'jumping links' in the perisylvian network, the auto-associative input through connections between the neurons of the developing circuits became strong enough to maintain activity for some time. As the individual circuits of the model correspond to word forms, this activity maintenance can be interpreted as verbal working memory, which may be important for building large vocabularies (Baddeley, 2003; Baddeley et al., 1998; Bishop et al., 1990). Therefore, the summarized work provides a candidate explanation why and how an evolutionary change in brain connectivity has led to a new skill, verbal working memory, and the ability to build large vocabularies in humans. The results also explain a range of previous observations, including the relationship between arcuate fascicle connectivity and word learning ability (Lopez-Barroso et al., 2013) and the relatively reduced ability of non-human primates to remember sounds (Fritz et al., 2005; Scott et al., 2012, 2014).

4.2. Symbol grounding mechanisms for semantic learning

The paragraphs above focused on the mechanistic basis of symbolic forms, of spoken words or written signs, still leaving it open how such forms bind to specific meanings in symbol learning within a neurobiologically plausible system. The connection between a word and the objects or actions it is used to speak about can be learned based on association. The Hebbian learning rule provides a biological basis of such associative semantic learning. Therefore, if the word "dog" is used in the context of perceptions of different dogs, the neurobiological links between neurons included in the mini-circuit of the word form may link up with neurons activated in visual (tactile, auditory, olfactory …) cortices

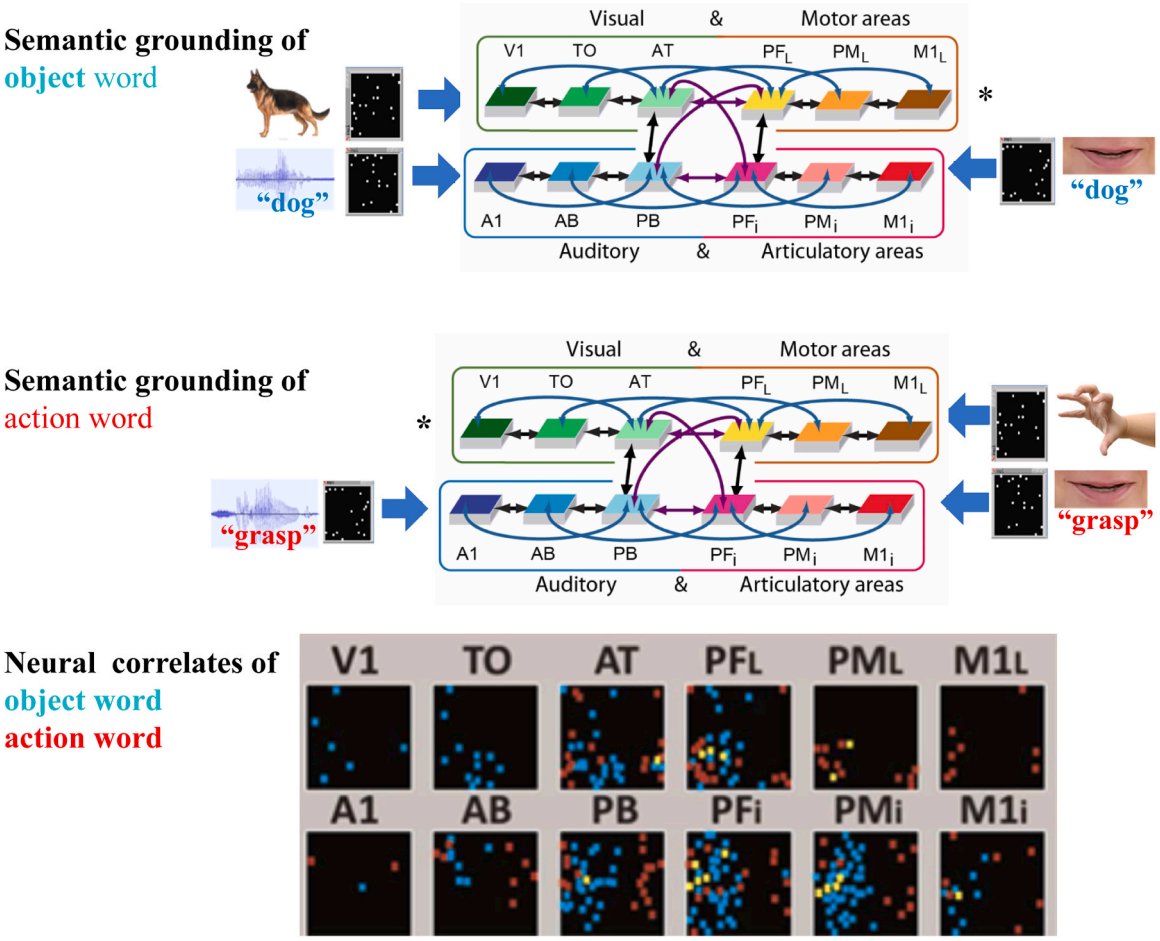


Fig. 3. Learning of symbol referent relationships in a brain-constrained model including 12 peri- and extrasylvian areas relevant for language and semantics. The model areas imitate the 6 areas in fronto-temporal perisylvian cortex of Fig. 2 plus 6 'extrasylvian' areas, including 3 dorsolateral frontal sites, lateral prefrontal, PF_L, lateral premotor, PM_L, and primary hand motor cortex, $M1_L$, and 3 temporo-occipital sites, anterior temporal cortex, AT, inferior-posterior temporo-occipital, TO, and occipital visual cortex, V1. The 12 area semantic model is used to simulate semantic learning of an object related word by co-activating neuron sets in perisylvian articulatory and auditory areas (imitating the production/perception of a spoken word form) plus neurons either in visual cortex (imitating the perception (or thinking) of a referent object, top panel) or in lateral motor cortex (imitating the execution of a referent action, middle panel). Bottom panel: As a result of co-processing word form and referent related information, distributed cell assembly circuits form which are spread out across model areas. The 12 model areas are spatially arranged as in the 12 boxes schematically displayed in the diagrams above. Individual colored dots index neurons of two different circuits, one simulating an action-related word (in red) and the other one an object word (in light blue). Yellow dots show neurons included in both word-related circuits. Most of the neuron members of these circuits are located in the connector hub areas of the model (areas PF_L, AT, PF_i, PB).
(a) Circuits of the object and the action word respectively include moderate but significant numbers of neurons in visual and dorsolateral motor areas (areas V1, TO, $M1_L$, PM_L), thus reflecting the different information sources of semantic grounding. (b) The figure includes artwork from (Tomasello et al., 2017).

(for illustration, see top panel of Fig. 3). This may interlink the word form mechanism with that of sensory perceptions of objects, thereby building a larger neuronal assembly now binding the word form to information about possible referent objects. A similar process may be triggered when action related words are used in the context of the learner's own actions: in this case, action related activity patterns in dorsolateral motor and prefrontal fields co-activate with word form circuits (see middle panel of Fig. 3, Pulvermüller, 2005; Pulvermüller and Fadiga, 2010).

Simulation studies using brain-constrained networks showed the formation of associative neuronal circuits (i.e., strongly interconnected sets of neurons, or cell assemblies) in a larger network imitating 12 perisylvian and extrasylvian areas relevant for processing words and their meaning (Garagnani and Pulvermüller, 2016; Tomasello et al., 2017, 2018). The 'extrasylvian' areas outside the 'language areas' and more distant from the sylvian fissure included primary and higher visual areas and anterior temporal lobe along with dorsolateral motor, premotor and prefrontal cortex. As a result of learning, strongly connected higher-order circuits developed, which interlinked linguistic circuits underpinning word forms with object and action related circuits. Consistent with previous neurocognitive theorizing (Braitenberg, 1978; Hebb, 1949; Pulvermüller, 1999), the linguistic and action/object related circuits in fact merged with each other as a consequence of learning. After learning, one single higher-order 'semantic circuit' for each stored symbol had formed, which integrates form and meaning information within one functional unit. Different versions of the brain-constrained semantic model varied the 'tightness' of model constraints, comparing, for example, a model composed of basic mean-field neurons with one using more sophisticated integrate-and-fire neurons (see Section 2), with comparable results (Tomasello et al., 2018). This consistency across model variants suggests a general semantic mechanism emerging in brain-constrained networks, which does not depend on the precise neuron model chosen.

While associative learning of information about word form and word referent may appear as a simple, almost trivial task, the model results provide important novel insights into the putative brain mechanisms underlying semantic processing. First, they provide a mathematically precise mechanistic model of conceptual and semantic grounding of symbols, the anchoring of symbolic meaning in entities and features of the world (see, for example, Barsalou, 2008; Cangelosi and Stramandinoli, 2018; Glenberg and Gallese, 2012; Harnad, 1990; Kiefer and Pulvermüller, 2012). Second, a host of neuropsychological data document so-called 'word category specific' deficits, whereby patients with brain lesions suffer from linguistic and semantic problems preferentially with specific word categories, for example action related verbs vs concrete object related nouns, or animal vs tool words (Damasio et al., 1996; Dreyer et al., 2020; Kemmerer, 2014; Shallice, 1988; Warrington and Shallice, 1984). In addition, metabolic and neurophysiological imaging studies have shown activation of different brain areas depending on and changing with the meaning of linguistic stimuli (Anderson et al., 2018; Binder and Desai, 2011; Kiefer and Pulvermüller, 2012; Martin, 2007; Pulvermüller, 2018b). Some of these activation and lesion results can be explained in detail by these simulation results (Chen et al., 2017; Garagnani and Pulvermüller, 2016; Ralph et al., 2017; Tomasello et al., 2017, 2018; Tomasello et al., 2019). Most importantly, the simulations indicate that the meaning carriers of the human brain are distributed circuits of neurons which, depending on their meaning, may be spread across different sets of cortical areas including primary and secondary modality-specific sensory and motor fields.

The diagram at the bottom of Fig. 3 shows examples of the contrast between the distributions of the cell assemblies of object and action related words as revealed by one of the model simulations (Tomasello et al., 2017). Note that the red dots showing individual neurons included in the cell assembly of one action-related word are relatively numerous in dorsolateral motor and premotor cortex (PM_L, $M1_L$), but not in visual areas in occipitotemporal cortex (V1, TO), whereas the blue dots indexing member neurons of object-word circuits are relatively more frequent in the visual areas. This difference in the distributions of cell assemblies is a direct consequence of the different patterns of sensorimotor activity during the simulated semantic grounding process. Therefore, this model explains why lesions in modality-preferential cortical areas frequently lead to category-specific deficits and why, in the intact brain, words of different semantic types activate specific modality-preferential areas. Similar theoretical proposals have been made previously (see, for example, Allport, 1985; Pulvermüller, 1999). However, the summarized simulation now demonstrates that these possibilities are also consistent with biological principles and observations manifest in the modelling constraints applied.

The simulation of the word types object vs action word just represents one example of a semantic category difference, where the categories also illustrate extreme cases although the semantic space is continuous. A more exhaustive simulation of semantic word types needs to cover additional referent-related information (see Binder and Desai, 2011; Borghi et al., 2022c; Harpaintner et al., 2020; Kiefer and Pulvermüller, 2012; Martin, 2016; Pulvermüller, 2018b). These not-yet-addressed aspects include: 1) information from other sensory modalities over and above the visual system, thus addressing auditory, tactile, olfactory and gustatory perception, 2) subtypes of modality specific information, as, for example, object related and spatial information in the visual and tactile domains or body part specific information in motor and tactile domains, 3) information from different modalities which is jointly relevant, as for example in the case of objects with clear action affordances such as tools or foods, where visual and other perceptual features (e.g., being red, round and well-tasting) are equally important as action features (e.g., graspability, manipulability) and 4) emotional load, which characterizes many constellations of perception and action related features. Taking these issues into account requires substantial extension of the 12-area semantic model shown in Fig. 3. More specifically, it calls for incorporation of additional areas (e.g., other modality-preferential cortices and emotion processing brain parts) and the more detailed modelling of already implemented ones (e.g., structuring the motor or auditory areas according to features such as somatotopy, retinotopy and tonotopy). Still, the addition of such important details will likely confirm the general conclusion from the 12 area model simulations, that the modality through which referential information enters or leaves the brain influences the cortical distribution of semantic circuits.

Apart from differences between the network correlates of semantic word types, Fig. 3 also shows similarities of cell assembly distributions. Neuron densities do not differ between word types in the perisylvian areas of the model and are also similar in some other areas outside perisylvian cortex – in particular in those distant from sensory or motor fields. The model areas in lateral prefrontal and anterior temporal cortex (PF_L, AT) are of this kind. Their central position in the entire network makes their neurons ideal for interlinking modality-specific information about word form and meaning and thus accumulating activity. Structurally, these areas are characterized by multiple links to other areas and therefore a high connectivity 'degree', which makes them so-called connector hub areas (Bertolero et al., 2018; Bullmore and Sporns, 2012; van den Heuvel and Sporns, 2013).

Fig. 3 (bottom panel) shows that the extrasylvian connector hub areas PF_L and AT include relatively more neurons of the semantic cell assembly circuits than the more peripheral modality-preferential areas. What explains this topographical feature? As a result of their high connectivity degree, connector hub areas receive a maximum of convergent activity from other areas. This results in comparatively enhanced activity of specific neurons in connector hub areas when symbols and their meaning are processed. In turn, this enhanced activity leads to particularly pronounced strengthening of mutual connections between the activated neurons (circuit consolidation) and to most efficient recruitment of new connected neurons into the cell assembly (Doursat and Bienenstock, 2007). In essence, the connectivity feature of

strong convergence on connector hub areas provides a candidate explanation for comparably strong cell assembly building and hence higher neuron densities of semantic cell assemblies there. Note, however, that this explanation does not rule out alternative or complementary accounts, for example related to the central position of these areas within the architecture.

Therefore, although, in the present simulations, activity patterns in visual, auditory and motor areas drive the model's semantic and symbolic learning processes (as regular stimulation is applied there), most of the neurons important for holding together the resultant semantic circuits are in the connector hubs. The underlying mechanism capitalizes on connectivity features and the resultant neural activity flow along with Hebbian learning and provides an explanation why there are so-called semantic hubs, that is, brain regions of general importance for conceptual and semantic processing. It is the connector hubs relevant for interlinking symbol form and meaning which, due to their great connectivity degree and convergence, become semantic hubs. Because of their high densities of semantic neurons, they do not only most strongly activate in different types of semantic processing, their lesion likewise leads to pronounced and general semantic deficits. Therefore, the model not only explains category specific semantic deficits after lesions in modality-preferential brain regions, it also accounts for a dominant and general semantic role of semantic hubs in frontal and temporal cortex (for evidence, see Binder and Desai, 2011; Kuhnke et al., 2020; Patterson et al., 2007; Ralph et al., 2017). Following the same line of argument, a future extension of the model to also incorporate inferior parietal areas may lead to an analogous explanation of the inferior parietal semantic hub (Binder and Desai, 2011). Note that the inferior parietal cortex has been found to be particularly relevant for processing spatial language and prepositions (Kemmerer et al., 2007; Shebani et al., 2021; Shebani et al., 2017; Tranel and Kemmerer, 2004).

Some current approaches to meaning processing in the human brain postulate one center or hub region thought to either house central semantic processes per se or bind symbolic form to meaning (for discussion, see Binder and Desai, 2011; Garagnani and Pulvermüller, 2016; Jackson et al., 2021; Pulvermüller, 2013, 2018b; Ralph et al., 2017; Tomasello et al., 2017). Interestingly, different theorists postulate their semantic hubs in different specific areas, for example the anterior inferior temporal lobe (e.g., Patterson et al., 2007) or the posterior superior temporal lobe (e.g., the 'lexical interface' of Hickok and Poeppel, 2007). The present neurobiologically constrained simulations suggest that all of these models are partly correct, as each of them highlights the semantic role of specific connector hub areas. However, in the brain-constrained model none of these areas has a special, let along unique, status. The 12-area semantic model already includes four areas where the density of semantic neurons is very high (AT, PB, PF_i, PF_L). A lesion to one of these may lead to a general semantic deficit. The model further implies that lesions affecting more than one of these connector hubs lead to most substantial semantic problems. In this context, it is noteworthy to consider that the lesions underlying semantic dementia typically start in the inferior anterior temporal lobe, but then expand to affect larger and larger parts of the temporal lobes, include superior temporal cortex (and thus area PB), and inferior prefrontal cortex (PF) (Hodges and Patterson, 2007; Mesulam, 2013). The prominence of the resultant semantic deficit may therefore be related to involvement of multiple connector and semantic hubs. In a nutshell, the summarized simulation studies are not only consistent with a broad range of facts known from neuropsychology and neuroimaging, they also explain the existence, and localization within the brain, of areas with a general and important role in meaning processing – the conceptual centers, semantic hubs, or lexical interfaces – along with the category specific semantic roles of modality-preferential areas important for semantic grounding – which are sometimes called 'semantic spokes'. Furthermore, the model also explains why the role of the former in semantics is substantial, whereas that of the latter is subtle (cf. Binder and Desai, 2011).

4.3. Fast mapping of form and meaning

Infants can learn aspects of the meaning of words rapidly. Already with one or a few presentations of a novel word form in the context of an object for which the infant has not learned a label previously, s/he can pick up the novel object-label relationship (Bion et al., 2013; Carey and Bartlett, 1978). This 'fast mapping' of form to meaning, or at least some aspect of meaning, is a remarkable feature of early language learning, which calls for biological explanation.

On first glance, the learning mechanisms most frequently applied in neural networks research appear as not ideally suited for such fast mapping, because of the gradual, incremental nature of the learning they exploit. With each learning event, each synapse's weight is only modified minimally, so that most simulations need hundreds or even thousands of learning events for storing knowledge in a network. Indeed, previous studies exploring symbol and semantic learning (for example, Rogers and McClelland, 2004; Ueno et al., 2011) used established error-gradient dependent learning algorithms and fully-distributed representations in feedforward networks, a strategy which, as discussed above, leaves room for improving biological plausibility. In addition, this approach seems unlikely to capture fast mapping.

Some studies applied biologically established Hebbian learning mechanisms to interlink object and word form representations which, respectively, were localized in different directly connected layers or network parts (Li et al., 2004; Li et al., 2007; Mayor and Plunkett, 2010). One of these studies (Mayor and Plunkett, 2010) found that application of a Hebbian learning rule including the associative component – i.e. the 'fire-together wire-together' term – led to reliable associative learning already after one or a few co-presentations of objects and their labels. However, several limitations apply to this study. First, the exclusive use of synaptic weight increase, under omission of synaptic weight reduction, captures only one aspect of biologically-realistic learning and ignores the possibility of weakening and 'unlearning' links due to non-concordant activations. Second, object and word representations were built in two layers or 'areas' and, in a second step, these representations were associated via direct between-layer connections. This two-layer network architecture provides a very much simplified picture of the cortical organization of semantic learning, in which multiple cortical areas are involved (Barsalou, 2008; Binder and Desai, 2011; Kiefer and Pulvermüller, 2012; Pulvermüller, 2018b; Shtyrov, 2011; Vasilyeva et al., 2019). With a realistic number of areas through which activity must travel to interlink symbolic form and meaning information, the mapping may become slower and less effective. Therefore, it is possible that a more realistic approach, where both Hebbian association and dissociation learning are applied in a neurocomputational model with multiple areas, leads to different results on the fast form-meaning mapping of symbols.

A recent study used the brain-constrained model with 12 simulated areas (see Fig. 3), to interlink phonological and conceptual circuits, which had first been set up in the perisylvian and extrasylvian parts of the networks (Constant et al., 2023). That fast mapping operates on previously acquired phonological and (pre-linguistic) conceptual knowledge is well established (MacNamara, 1972; Mayor and Plunkett, 2010; Schyns, 1991). Hebbian learning, including both its associative and dissociative components, was implemented and learning dynamics traced across 100 learning events per word-instance pairing, by probing, after each learning event, whether semantic information was retrieved following network stimulation with the word form pattern only. Results showed that already after the very first learning event, significant semantic information was retrieved for some of the word forms. Already after 10 learning events, there was evidence for learning of ca 70% of the word-meaning mappings; after 20-30, network performance was already close to 100%. Learning efficiency was speeded by increasing the network's 'attention level', which was implemented by reducing the amount of competition in the network. Overall, the study shows that

realistic Hebbian learning in a network mimicking the human cortical machinery for language and concept processing can explain fast semantic mapping (Constant et al., 2023).

5. Learning and concept formation induced by proper names and category terms

The previous section focused on semantic learning in a simple and straightforward sense. Spoken words, written symbols – regardless of these are spoken words, written signs or hand gestures – become meaningful by interlinking them with information about entities in the world, including objects and actions (Frege, 1892). That inter-individually accessible links with the real world, which semantically 'ground' symbols, are established is a necessary condition for guaranteeing that different people speak about the same issue when using the same signs (Harnad, 1990). This condition is violated by approaches construing the semantic link between form and symbol as one immanent to mental processes of each individual per se (see, for example, de Saussure, 1916; Jackendoff, 2002); in this case, it remains still to be explained how the interpretability of signs is similar in different individuals speaking the same language (Alston, 1964; Baker and Hacker, 2009; Wittgenstein, 1953). Even if they learn the same symbol forms and are a priori equipped with the same concepts, an explanation is necessary how they can link each symbol with the same target concept, but not a different one. A shared practice of using the same word to speak about the same real-world entities explains symbolic interpretability (Harnad, 1990), at least for some types of symbols.

However, bridging between symbols and real-world entities is only one out of many functions language can carry. And, upon closer examination, most vocabulary elements do not 'just label' single objects. Rather, they may be used to speak about quite different 'things'. And the class or category of entities they can be used to speak about may not be sharply defined and therefore may allow elements for which category inclusion and 'label application' is disputable (Löbner, 2013). The categories themselves may be small and easy to describe, or they may be large and include many atypical members or, in the extreme, be entirely heterogeneous. For the latter case, think of words such as "beauty" or "justice", which can be appropriately applied to many different entities or instances. In the simplest case, the symbol is specific to one single object. For such 'proper names', the semantic link between object and word can be built straightforwardly by an associative 1:1 mapping. In contrast, more sophisticated mechanisms are required for the mapping between a category term and the range of entities it can be applied to (Pulvermüller, 2018c; Westermann and Mareschal, 2014). In addition, the learning of abstract terms and meanings may necessitate processes different from mere association (Dove, 2016; Machery, 2016).

What is the neurobiological basis of gradually more complex semantic links? In addressing this question, the focus will first be on category terms in comparison with proper names, before (Section 5), later-on, the discussion will expand to also include abstract concepts and words (see Section 6). In contrast to proper names of individual entities, for example persons, category terms cannot be construed as 'labels' for objects or other entities. They rather seem to be tools for distinguishing entities that fall into a given category from those that do not. Such categorization may have implications not only for language, but for a broader range of cognitive operations, including thought (Kemmerer, 2022; Lupyan et al., 2020; Majid et al., 2004; Majid et al., 2018; Westermann and Mareschal, 2014).

When contrasting learning mechanisms for proper names and category terms, it will immediately become clear that a mechanistic neurobiological model of language learning has implications not only for language immanent, linguistic knowledge, but affects attention and perception mechanisms too. Therefore, this present section will focus on the effect of language and symbol learning on attention. Originally, such an effect was seen as general, insofar as 'labels' were found to boost attention to their related reference object. However, recent developmental experimental research reveals a quite fine-grained and sophisticated picture according to which different types of verbal expressions bring about different effects of attention modulation, directing the attentional mechanism to different features of reference objects. A brain-constrained model will be developed to explain why different features are being attended to when processing the referents of proper names specific to an object and category terms applicable to a whole class.

5.1. Object-related symbols as attention enhancers

There is important evidence for influences of language learning on other cognitive domains. In particular, developmental investigations in young children show that learning a word for an object leads to attention increase and better memory for the respective object. Baldwin and Markman introduced infants to novel toys, either with or without using a different label for each toy (Baldwin and Markman, 1989). In a subsequent phase, the infants showed a looking preference for the labeled objects. This is evidence that the children attended relatively more to the objects after encountering them in the label context and that they remembered these objects, or at least some of their features, relatively better. Because infants attend more to objects for which they know or are given a verbal symbol, language has been proposed to function as an 'attention enhancer' (e.g., Sloutsky and Robinson, 2008). The attention and memory enhancing role of language is also reflected in neurophysiological recordings. For example, brain responses to objects vary depending on whether 12-year-old infants know 'names' for these objects (Gliga et al., 2010). Concordantly, responses to words were found to increase with the number of times these were linked with concrete meaning (Aleksandrov et al., 2020).

Different theories have been proposed to provide accounts for an attention-enhancing role of language, including behavioral statistical approaches claiming that this property is a consequence of co-occurring verbal and non-verbal information and the Hebbian associative learning rule (see Mayor and Plunkett, 2010; Sloutsky et al., 2017). In the brain-constrained semantic model described above (Fig. 3), the attention enhancing effect can be explained by flow of activity from the perisylvian word form circuit to the associated neuron set activated by the perceived object or action (Chen et al., 2017; Garagnani et al., 2016; Tomasello et al., 2017). In sharp contrast with these models, cognitive theories claim that an explanation in terms of association is insufficient because even very young children apply theory-driven constructs, including semantic and syntactic representations, already when they learn their first words and phrases (Gleitman, 1990; LaTourrette and Waxman, 2020; Lidz and Gleitman, 2004; Perszyk and Waxman, 2018; Waxman and Gelman, 2009). In this perspective, already young infants are viewed as 'theorists' who build representations of objects and categories characterized by abstract features and draw inferences about these representations and the entities to which they refer.

Whether infants build object representations (in the sense of cognitive theory), rather than map sensory experiences, has been addressed experimentally. For example, Preissler and Carey co-presented infants of 1.5–2 years of age with object pictures and pseudowords and tested whether these infants subsequently matched the pseudowords with similar photographs of the same objects, or rather with three-dimensional model objects (Preissler and Carey, 2005). Infants selected either the objects alone or both objects and pictures, but not pictures alone, although these latter items were visually most similar to the pictures used during learning. The authors argue that, because perceptual similarity between two near-identical object pictures is greater than that between a three-dimensional object and a 2-dimensional picture thereof, the infants should have chosen pictures only if similarity had been critical. They conclude that the infants' selections were guided by 3D object representations rather than just by visual similarities between the stimuli. These results were interpreted as evidence that young children build neuronal machineries for objects and

concepts that abstract away from individual perceptual instances and are specific to whole objects or even categories. A suggested further conclusion is that children are not just associators, but rather theorists who a priori know that it is objects that verbal labels interlink with. On the other hand, as infants receive ample information relevant for building 3D object representations and the emergence of such representations can itself be seen as a consequence of associative learning, the partial preference for whole objects seems consistent with, and explainable by, associative learning models (Sloutsky, 2009). In conclusion, this research appears to be open to different interpretations and cannot decide between associative and cognitive-theorist approaches.

5.2. Different attention effects of category terms and proper names

Over and above a role of language in generally enhancing attention to and memory for objects, cognitive scientists claimed that language also facilitates the formation of conceptual categories (Brown, 1958; Perszyk and Waxman, 2018). One paradigm with which this issue was addressed uses different perceptually similar objects (e.g. toys or animals). After encountering different similar objects from the same category always co-presented with the same word form, a probe pair of images is presented, showing one object again from the same category (another toy or animal) and a quite dissimilar object from a different category (e.g., a fruit). Infants, even as young as 12 months, show a looking preference for the novel-category object. This result is interpreted as evidence that the infants have formed a conceptual category in their minds and are therefore surprised by the occurrence of a member of a different category, and consequently look at it. If no spoken word form is co-presented with the similar objects, the latter effect was absent, thus suggesting that category formation is boosted by the presence of the verbal form co-occurring with the category members. Similarly, no category building effect is reported if the co-presented items are either a word with very wide use (e.g. 'look', 'there'), a single non-verbal sign such as a tone sequence, or different verbal labels each specific for one object/picture (Balaban and Waxman, 1997; Ferguson et al., 2015; Fulkerson and Waxman, 2007; Perszyk and Waxman, 2018; Waxman and Braun, 2005).

These data support the role of language as facilitator of conceptual and semantic category formation. Given that many conceptual categories are characterized by perceptual features shared by all category members, cognitive scientists concluded that these shared features receive relatively more attention due to the fact that the category-congruent patterns have been 'labeled' with the same word/pseudo-word during learning (Althaus and Mareschal, 2014; Gelman and Waxman, 2009; Perszyk and Waxman, 2018; Waxman and Markow, 1995).

However, if labels indeed drive attention towards shared features of objects, the following finding appears most peculiar and surprising: If similar objects sharing visual features (e.g., different toys) are labeled by specific terms, so that each object has its specific 'proper name' (instead of a 'category term' for the entire set), young children tend to focus their attention on the features distinguishing these objects from each other (rather than on the shared ones). This focusing on specific object features is evident from a recent study where infants, again already at the age of 12 months, were shown similar toy pictures paired with pseudowords and thereafter were shown the pictures again, each with a novel toy. The infants tended to look at the novel toys. This finding cannot be explained by category formation, as all stimulus pictures contained members of the same perceptually similar category, toys. Therefore, the authors argue that "hearing distinct names applied to the same objects focuses infants on the uniqueness of each object" (LaTourrette and Waxman Sandra, 2020).

These results demonstrate intriguing features of early-life language-attention interactions which are not easily captured by general 'attention enhancement' or 'capturing'. However, attributing category formation to consistent or inconsistent labeling of similar objects per se can hardly provide an explanation for these surprising phenomena. How can a label 'build' a category? What might guide the 'child as a theorist' (see LaTourrette and Waxman Sandra, 2020; Perszyk and Waxman, 2018; Waxman and Gelman, 2009) to follow different feature attention strategies when encountering objects with consistent and inconsistent 'labels', i.e. proper names and category terms? Which mechanisms may underlie the finding that children, who encounter the same word form together with similar objects that share visual features, tend to attend to these shared features rather than to the specific features distinguishing these objects between each other? Likewise, it needs to be explained why proper names for specific objects results in better memory for these individual objects and attend to their specific features, and why 'category building' is absent (or reduced) in this case.

It is clear that a mechanistic account of these diverse and sophisticated attention effects of symbols is necessary. It is difficult to see how associative links between word forms and objects could provide such an account. What distinguishes the semantic links of object-specific and category-general 'labels', so that their different attention-driving roles can be explained? Such an explanation requires a closer look at the mechanistic neuronal basis of category formation and referential-semantic learning.

5.3. A neurobiological perspective on proper name and category term formation

In an associative learning framework, the co-occurrence of objects and labels should strengthen the links between the neural correlates of object and label knowledge (see, for example, Lupyan, 2012a; Lupyan, 2012b; Westermann and Mareschal, 2014; Westermann et al., 2006). So, how is it possible to explain the labels' differential direction of attention, either to specific or general shared features of objects, dependent on the specificity of symbol use? Here is a neurobiological account emphasizing the role of dissociation learning in sharpening the specificity of the links between object-related and symbolic information and resulting in different mechanisms for object-specific proper names and category-general terms.

The explanation rests on the assumption that, when word forms and their related referent objects or actions are learned, not whole-entity mechanisms are relevant (as the representation, or within-brain correlate, of an object or its photographic image) but that, instead, sets of features are stored. That feature-specific neurons are common in the brain has first been shown for visual perception, where neurons maximally responsive to features at different levels of complexity have been demonstrated to exist at different levels of the perceptual processing hierarchy (Hubel, 1995; Hubel and Wiesel, 1977). Feature-specific neurons have also been shown for other modalities, including, for example, sounds (Mesgarani et al., 2014; Steinschneider et al., 2003; Yi et al., 2019). Likewise, the motor system houses feature-specific neurons whose activity indexes, for example, whether a to-be-pronounced phoneme is labial, alveolar or voiced (Bouchard et al., 2013). Note that this statement is compatible with the existence of neurons for very specific features, or even object-specific ones in the extreme, at the highest levels of the hierarchy (see Barlow, 1972; Fuster, 1995; Palm, 1982; Quiroga et al., 2008). Therefore, neuroscience evidence is consistent with the assumption that a given perceived object or to-be-performed action is reflected in the activation of a range of neurons at least some of which are specific to perceptual and motor features of different complexity.

Whereas the assumption of basic perceptual and action-related features (such as MOVES, IS HORIZONTALLY ORIENTED, IS EDGY) is straightforward and well-established in brain research (see, for example, Hubel, 1995), the higher-level semantic features assumed by semantic theories (e.g., ANIMATE, MALE/FEMALE, YOUNG/OLD etc.) have a more complex relationship to sensorimotor features. One pathway to explore the putative mechanistic basis of such higher-level

perceptual-semantic features is to use artificial neural networks, for example deep neural networks applied in classification tasks (see Kietzmann et al., 2019b) or auto-encoder networks with a central layer of a few latent units (e.g. Higgins et al., 2016; Johnston and Fusi, 2023). Interestingly, neuronal activity patterns in different layers of these networks sometimes show specificity to lower- and higher-level features comparable with that of different cortical areas (Kriegeskorte and Kievit, 2013; Liu et al., 2013). In particular, the latent units in the central layer of auto-encoders can show stimulus specificity consistent with high-level perceptual-semantic features and also similar to neurons in different cortical areas (Bernardi et al., 2020; Higgins et al., 2021; Ito et al., 2022). This neurobiological and neurocomputational research may reveal mechanisms for higher-level feature extraction and generalization, although relating an auto-encoder network's algorithm to biology may still appear as a challenge. The results provide further motivation for assuming that neural representations are built from lower- and higher-level perceptual features that can also play a conceptual or semantic role.

In this perspective, auditory, visual, motor and other modality-preferential systems extract and store perceptual and action-related features at different levels and combinations thereof, from which symbols, concepts and meanings are built compositionally (Pulvermüller, 1999, 2018a; c). The neuronal set for a word form would therefore include a structured set of neuronal elements related to articulatory and acoustic phonetic and phonological features, and that of a concept a set of perceptual and action-related (including interaction-related) feature neurons. Circuits corresponding to phonologically similar word forms share phonetic feature neurons and network correlates of similar objects and actions share neurons for perceptual and action-related features.

If objects fall into the same conceptual and semantic category, they typically share perceptual and action-related features characterizing the category. For example, dogs typically have 4 legs, 2 eyes, fur and tend to bark, chase rabbits and fetch sticks thrown for them by humans. Given these features have neuronal correlates, not necessarily one single neuron per feature, but likely several, possibly differing between each other in their response characteristics. If two objects are similar, they share features and, thus, the neuron sets they activate will overlap in some of their feature neurons. Hence, these 'overlap neurons' shared between object representations become 'semantic neurons' and therefore the mechanistic basis of a concrete semantic category. Still, other neurons responsive to individual objects may not be shared and thus remain perceptual feature neurons as they correspond to more specific or even 'idiosyncratic' features of these objects. Fig. 4a illustrates the putative neuronal basis of a conceptual/semantic category by showing 3 overlapping neuronal sets, with semantic neurons in the overlap area (indexing common features, e.g., 'has 2 eyes', 'barks' etc.) and perceptual neurons in the remaining, specific parts of the sets (indexing instance specific ones, e.g., 'is orange-brown colored'). For illustration and clarity, only object and category specific neurons are shown – although the sets may include other neurons too, e.g., neurons responsive to a broad range of category members, but not strictly all of them ("has fur" applies to most, but not all dogs), or neurons indexing a conjunction of aforementioned features).

Now consider the case where one novel word form or symbol is presented together with each of the members of a semantic category – as in the experiments on category formation mentioned above. The shared semantic neurons will be active each time the word form appears. Therefore, any connections that link word form circuit and the shared neurons will strengthen each time one of the objects appears together with the word (solid lines in Fig. 4b). However, if we assume that n category members are available and all members appear equally often, the links between each of the specific parts of the object related circuits (which include specific/idiosyncratic perceptual feature neurons) will only strengthen in 1/n of the cases of object-word pairing (n = 3 in the illustrated example, resulting in 1/3). This however only takes into account associative learning; the dissociation term of the Hebbian learning rule adds to the difference: In addition, each of the links between the object-specific parts of the sets and the word form circuit will weaken whenever an instance-specific feature neuron is silent while the symbol form circuit activates – which, in the illustrated case always happens for the two instance circuits not active when the third instance is co-presented with the word form (in (n-1)/n cases; resulting in 2/3 for n = 3). Therefore, the co-presentation of category terms with individual referent objects connects the word form circuit with semantic neurons shared by all category member objects, which cognitively relate to the knowledge about the shared features of a conceptual and semantic category. In contrast, the information about object specific features is not being strengthened, and may even be delinked from the category term. This model implies that category terms induce most activation in the neurons of the semantic overlap. At the cognitive level, the mechanism explains why category terms direct attention to shared features of the category members, rather than to their specific features.

The opposite dynamics is induced by the formation of links between objects and their specific labels or proper names. In this case, the neurons of the semantic overlap strengthen their links with the word form circuits whenever a given object appears together with its proper name (in 1/n of the cases, 1/3 in the example). When an object-specific term appears with a different object, dissociation learning and synaptic weakening are effective for all neurons in the semantic overlap (in (n-1)/n of the cases, 2/3 in the example). In contrast, the idiosyncratic or

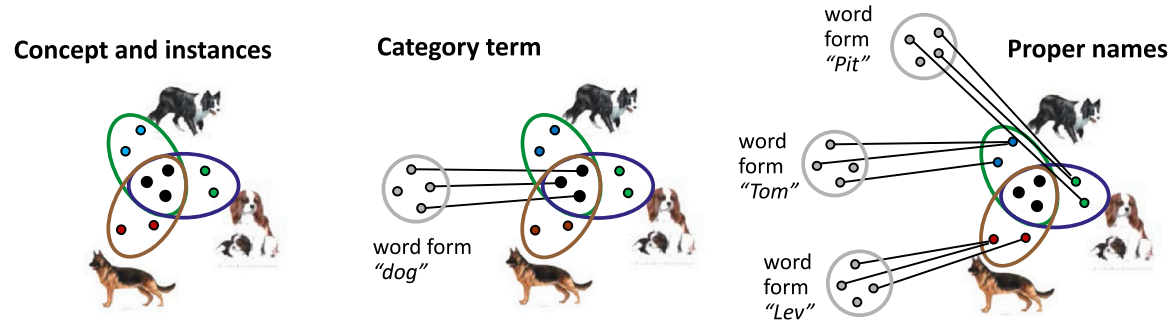


Fig. 4. Model of neurobiological mechanisms underlying the processing of objects, concepts, category terms and proper names. Left panel: Individual instances of objects are processed by activating sets of strongly connected neurons responsive to features of these objects (dots included in the ovals). A concrete conceptual category of objects is processed by activating neurons that index common, shared features of the category (black dots in the intersection area). Middle panel: The link between a category term and the concrete conceptual category of objects it relates to is mechanistically implemented by strong connections between the word form circuit (on the left) and the semantic overlap of neurons indexing shared semantic features of the category (black dots). Right panel: The links between proper names and the specific objects they relate to are mechanistically implemented by strong connections between the word form circuits (on the left) and those neurons of the object related circuit that are specific to the object (colored dots). As explained in text, these different connectivity patterns are implicated by Hebbian learning of associations and dissociations of linguistic, perceptual and conceptual features.

object-specific neurons show the reverse pattern, association with the co-occurring specific label (in 1/n of the pairings; 1/3 in the example), but dissociation from the other word form circuits (again in 1/n of the cases; 1/3). Assuming a larger learning constant for association than dissociation learning, the only neurons with strong links to the word form circuit of the proper names will be the object-specific ones (solid lines in Fig. 4c). This mechanism explains why proper names direct neuronal activity and attention to the specific feature neurons and the specific perceptual and cognitive features of their referent objects.

These general predictions were confirmed by recent model simulations of proper names and category terms. Learning of the latter led to circuits characterized by a predominance of shared semantic neurons across the central areas of the 12-area semantic architecture of Fig. 3, which by far outnumbered the shared-feature neurons of matched proper-name circuits. In contrast, the emergent neuronal assemblies of proper names included significantly more instance-specific neurons than those of category terms (Nguyen et al., 2023). The attention-driving role of proper names and category terms towards unique vs shared features of referent objects is therefore reflected and manifest in the relative numbers of instance-specific vs semantic neurons in the model simulations. This result demonstrates that Hebbian learning in a brain-like architecture can explain the attention-driving role symbols and language exert on object perception.

In summary, a biologically motivated model at the mechanistic neuronal level including Hebbian association and dissociation learning explains the relatively stronger linkage of proper names to object-specific features of their referents along with the fact that category terms attract attention to the shared features of the semantic category. These mechanisms may lay the ground for the complex cognitive activities of young infants, which are open to descriptions in terms of their theorizing and drawing conclusions about referents, concepts and their features.

6. Concrete vs abstract concepts and meaning

Similar to the previous one, this section will focus on the learning of different kinds of concepts and symbolic meanings, and on modelling the mechanistic basis of these in brain-like neural networks. Whereas Section 5 discussed specific vs general symbols, differences between concrete and abstract concepts and symbols will now be in focus.

6.1. Grounding of abstract terms: indirect or direct?

Abstract concepts and the semantics of abstract words present a challenge to many semantic theories (Barsalou et al., 2018; Borghi et al., 2020; Dove, 2009, 2016; Fischer et al., 2021; Glenberg, 2021; Glenberg and Robertson, 2000; Vigliocco et al., 2014). Models based on semantic features describe concrete concepts by concrete features such as GREEN, FOUR-LEGGED, LONG and SNAPPY; which can be related to, and grounded in, perceptual features and action-related ones (Katz and Fodor, 1963; Löbner, 2013). However, when it comes to abstract terms, such as “democracy” or “beauty”, semantic theorists recur to abstract features such as DEMOCRATIC or BEAUTIFUL (see, for example, Mahon and Caramazza, 2008), a strategy just moving the need for explanation from the level of concepts to that of features. Therefore, although this strategy offers economic descriptions of the semantics of huge vocabularies with a limited set of features, it does not address the question of how concepts relate to the real world in which children learn word meanings in the context of experiences. And even if one is inclined to hold that concepts are given to humans a priori, there would be need to connect concrete real-life events including objects and actions with the presumed internal *a-priori* entities by learning. The semantic feature approach does not offer explanations for such conceptual learning and grounding as far as abstract semantic features and hence abstract concepts are concerned.

Other models treat concrete and abstract concepts in the very same way, suggesting that any apparent differences might not be fundamental. *Distributional semantic models* define concepts and meaning based on information about the contexts in which words expressing concepts appear (Landauer and Dumais, 1997; Schwanenflugel et al., 1988). This strategy rests on the assumption that conceptual and semantic similarity of symbols is manifest in the co-occurrence of these symbols in texts. Semantic vectors describing such patterns of co-occurrence are useful for describing semantic relationships between symbols. However, to extract meaning from symbols and their contexts, it is not sufficient to describe semantic relationships between symbols. As mentioned above, it is also necessary to clarify what the symbols are used to communicate about (see Searle, 1980; Searle, 1984). Such semantic grounding is not covered by an account defining symbolic meaning in terms of other co-occurring symbols, and, therefore, distributional semantics alone cannot suffice to explain concepts, as it runs into the symbol grounding problem (Harnad, 1990; Searle, 1980). At least some concepts and symbols require conceptual ‘grounding’ in specific information about entities in the world, that is, in concept-related objects, actions or their features. Only then can distributional learning work via contextual transfer of conceptual information – which has also been called indirect grounding or ‘symbolic theft’ (Cangelosi et al., 2000, 2002; Cangelosi and Harnad, 2001). A minimum of ca. 10–20% of the words of a vocabulary may need to be directly grounded in entities in the world, so as to allow for indirect grounding based on context-based distributional learning (for discussion, see Blondin-Massé et al., 2013; Vincent-Lamarre et al., 2016).

One proposal is that the meaning of concrete symbols can be learned by direct grounding, for example co-presentation of a toy and its name to an infant, but that abstract terms need to be grounded indirectly, by use in linguistic contexts, together with other directly grounded concrete ones (see Borghi et al., 2019; Dove, 2009, 2010). Indirect grounding of abstract concepts was probed in a neurocomputational study. A neural network modeled the acquisition of abstract category terms, such as “use” or “make”, by frequent co-occurrence with previously grounded more specific expressions designating members of the category (hammer or fork use, hole or noise making, see Cangelosi and Stramandinoli, 2018; Stramandinoli et al., 2017). However, please note that these examples address one specific sub-type of abstract symbol, which can be used to speak about a variety of actions, whereby related more specific action expressions are already grounded. For other abstract terms (including action, object or property-related ones), such related, more specific and already grounded expressions may not be available. For example, the words “beauty”, ”truth” or “democracy” can only be grounded indirectly, if symbols with reasonably similar meaning are already known and grounded semantically. To ground the abstract term “truth” indirectly in contexts including “belief”, “think” and “idea”, at least some of the equally abstract context words need to be known. However, recent corpus studies suggest that abstract words primarily co-occur with other abstract ones (Lenci et al., 2018; Naumann et al., 2018), thus casting doubt on the feasibility of their indirect context-based grounding. It seems unavoidable to explore the possibility of directly grounding at least some of the highly abstract words in objects, actions and their features.

But is it possible at all to directly ground abstract concepts and meanings in real world events? Some theorists deny this, based on principal considerations. For example, it has even been argued that abstract concepts have a different ontological status than concrete ones insofar as “abstract entities are not in spacetime whereas concrete entities are” (Hale, 1988), a position seemingly excluding the possibility of their direct semantic grounding. However, it is undeniable that both abstract and concrete concepts are in fact concepts and, therefore, in one sense, not in the world, where space and time apply, but rather in the mind. Nevertheless, both concrete and abstract symbols need to be applied in real life to make claims and to confirm or reject them. After all, whether the statement “this is democracy” (or “democratic”) is correctly applied viz the current practice of voting in the market place of

Bern or at the presidential election in Uganda is an empirical issue. Note that this question is comparable to (although more complex than) that of whether a given animal can be called "a dog". Hence, as statements including both abstract and concrete terms need to potentially undergo verification or falsification, there need to be *criteria* for matching concepts with entities in the world and their features (Locke, 1909/1847). Psychological experiments where subjects are asked to list their situational associations for concrete and abstract concepts further confirm that both are intrinsically linked to background situational information and that these links are central to their meaning (Barsalou and Wiemer-Hastings, 2005). Therefore, it is obvious that at least some abstract terms, like concrete ones, need to be grounded and that this is possible by relating them to real world instances of the concept – e.g., observable democratic practices (for detailed examples of directly grounding concepts, see Fischer et al., 2021; Glenberg, 2021; Pulvermüller, 2018a). In this perspective, direct grounding is required for at least some members of each family of abstract terms with similar meaning.

6.2. What makes concepts abstract/concrete?: cognitive theories

In search of specific differences between concrete and abstract concepts and their grounding in 'world relationships', psychologists and linguists have highlighted several features. The dual coding theory postulates that abstract concepts and words are represented in a verbal system, whereas only concrete ones are represented by both verbal and imagery codes (Paivio, 1971, 1991). However, given the situational links of abstract concepts documented empirically (Barsalou and Wiemer-Hastings, 2005), it appears problematic to exclude an imagery code for abstract entities. In an alternative perspective, the difference rather lies in qualitatively different imaginistic codes for the two concept types, with concrete concepts offering relatively more sensory and motor associations and abstract terms more affective-emotional ones instead (Kousta et al., 2011; Paivio, 2013). However, this position seems to be driven by concepts that are abstract because they relate to emotional states, for example JOY, SORROW, LOVE and AGONY. Abstract mental concepts, such as LOGIC, CAUSALITY, NUMEROCITY and PROOF, are not rated high on emotional semantics (see, for example, Dreyer and Pulvermüller, 2018). Admittedly, students may classify the subject of logic as something positive or negative, but such judgements are variable across subjects and therefore not relevant for the meaning of the related word. Abstract terms related to logic, mathematics, numbers or aspects of cognition do not seem to be generally emotion-laden, although some words from these groups certainly are ("minus", "thirteen", "dream"). In fact, a specific sub-set of abstract words, those used to speak about emotions themselves, is primarily characterized by affective-emotional semantics, but not the abstract category as a whole (Dreyer and Pulvermüller, 2018; Moseley et al., 2012; Pulvermüller, 2018a). It seems that some of these abstract emotion words are learnt in contexts where the internal emotional state is manifest in the behavior and actions of the infant, thus providing the necessary criteria for correct application of the 'internal state' symbol by the adult (Gebauer, 2017; Holodynski, 2017; Wittgenstein, 1953). This research suggests that abstract emotion words are a possibly somewhat atypical subtype of abstract words, which closely resembles action related words, as the semantic grounding of both is characterized by learning in rather specific action contexts (Glenberg and Robertson, 2000; Moseley et al., 2012; Moseley and Pulvermüller, 2018).

A related perspective views external and internal attributes as relatively more crucial for concrete and abstract concepts, respectively, based on the fact that study participants tend to describe concrete concepts (e.g., BIRD) by using concrete action and perception related words ("beach", "fly", "food"), whereas for describing abstract concepts (e.g., TRUE), more abstract 'introspective' terms are applied ("belief", "think", "idea", Barsalou and Wiemer-Hastings, 2005; Borghi et al., 2022a; Borghi et al., 2022b). For this proposal too, a range of abstract concepts do not fully fit the pattern, for example abstract attributes applicable to concrete objects or actions, including BEAUTIFUL, RISKY and PROVOCATIVE. Although these relate to internal states (pleasure, uncertainty etc.), the corresponding word meanings can be explained just by describing or pointing to real-world examples. All these perspectives on abstractness or abstraction – the dual coding, emotion and introspection approaches – interlink different domains of semantic content (imagistic/linguistic, sensorimotor/emotional, external/-internal) with concrete and abstract concepts, but do not see a principal structural difference between them.

A structural description of the difference between concrete and abstract concepts goes back to the observation of a feature called *family resemblance* (Baker and Hacker, 2009; Wittgenstein, 1953). As mentioned above, the classic approach to category structure, that a distinctive set of semantic features are shared between the members of a category, is already not fully sufficient for large categories, such as BIRD, where some category members indeed lack the core feature of FLYING (e.g., PENGUIN), and certainly fails for more abstract large categories like GAME, where features such as GROUP ACTIVITY, PLEASANT and COMPETITIVE apply to subgroups of instantiations, but not to the entire set of activities falling under the term. There is a tension amongst semantic frameworks, where one fraction advocates, in spite of these counter-examples, the classic idea of common semantic features defining a concept and the other the general applicability of family resemblance (see, for example, Baker and Hacker, 2009; Löbner, 2013; Rosch and Mervis, 1975). A recent proposal is to apply the family resemblance feature for distinguishing abstract from concrete concepts and for characterizing a gradual abstract-concrete dimension (Pulvermüller, 2013, 2018a).

According to this approach, concrete concepts (and meanings) share a set of common semantic features, whereas abstract ones do not. Instead, each abstract concept is characterized by partial feature sharing, so that all semantic features are common to just a subset of instantiations falling into a given category. As discussed elsewhere (Henningsen-Schomers et al., 2023; Henningsen-Schomers and Pulvermüller, 2022; Pulvermüller, 2018a; c), this position captures a broad range of concrete vs abstract concepts and meanings. Furthermore, family resemblance accommodates the more basic structural property of abstractness (Langland-Hassan et al., 2021; Löhr, 2022; Lupyan and Mirman, 2013; Sloutsky, 2010), that instances of abstract concepts share less commonalities among each other, and lack shared features altogether, and provides a more specific description. However, not all symbols normally classified as abstract show the family resemblance feature. Large category terms do show it only to a small degree (e.g., most BIRDS share typical BIRD features), as only a limited subset of a category lacks semantic core features (PENGUIN, OSTRICH) and therefore may appear as 'non-prototypical'. In addition, abstract emotion words are rated as highly abstract although they share semantic features grounded in action execution (e.g., HAPPINESS grounded in emotion expression by smiling, laughing etc.). However, the full vs only partial feature sharing difference describes a structural difference between many concrete and abstract concepts and terms and also allows for a gradual continuous transition from extreme (structurally) abstract to fully concrete semantics. This is important for modelling the many concepts and word meanings half-between the extremes (for discussion, see Pulvermüller, 2018a; Pusch et al., 2023). Furthermore, the full vs partial feature sharing model captures the contextual variability and flexibility of the meaning of abstract symbols, which was previously noted by researchers in the field (see, for example, Barsalou and Wiemer-Hastings, 2005; Borghi et al., 2022b).

Fig. 5 schematically illustrates the concreteness/abstractness difference in terms of full vs. partial semantic feature overlap. In this display, each small circle represents an individual neuronal element thought to carry one specific perceptual or action-related feature activated by one or more *instances* of a concept, the real-world entities that 'fall under' the concept and the sensory and motor *patterns* of neuronal activation informing about them. One can classify these neurons into unique

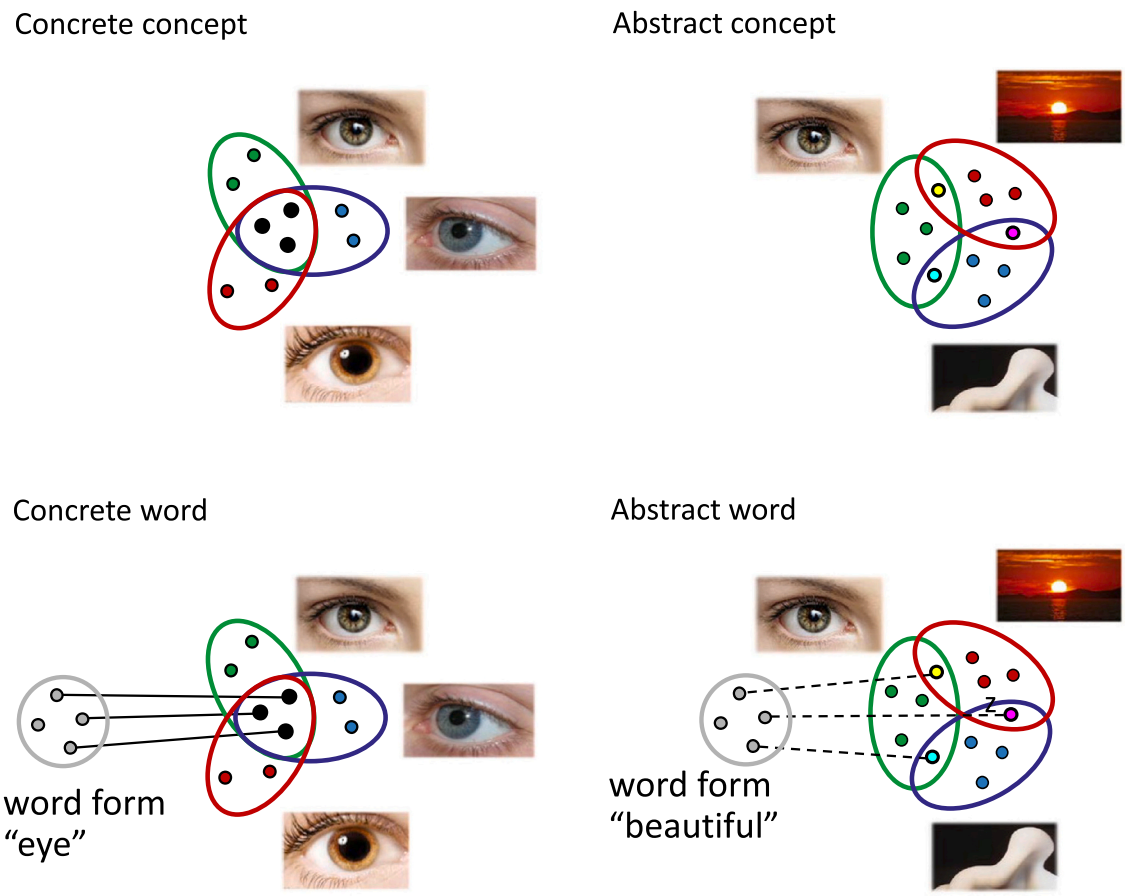


Fig. 5. Model of neurobiological mechanisms underlying the processing of concrete and abstract concepts and that of concrete and abstract words. The panels on the left show concrete items, whereas those on the right show abstract ones. Models for concepts appear at the top, those for meaningful words at the bottom. The semantic overlap structure and symbol linkage of concrete concepts and symbols corresponds to those in Fig. 4. A shared overlap of semantic feature neurons (black dots in the intersection areas) characterizes concrete conceptual and symbolic mechanisms. Abstract concepts and symbols typically lack semantic features shared by all instances that fall under the concept. Instead, a pattern of family resemblance and only partial feature sharing is common. The model predicts that concrete but not abstract concepts can be learnt from experiencing conceptual instances, due to high vs. low correlation of semantic neuron activation. A further model prediction is that concordant activation of symbols is beneficial for concrete concept formation, but essential and possibly even necessary for that of abstract concepts. For explanation, see text.
(The bottom panels are reprinted with permission from Pulvermüller, 2013).

specific neurons (responsive to only one instance of the category) and shared semantic neurons (activated by several (or at least more than one) instance(s)). Circuits for concrete concepts are characterized by a core set of semantic neurons shared by all (or most) instances, whereas the neuronal correlates of abstract concepts include no such or a minimal core set insufficient to define the category. Instead, these are characterized by semantic feature neurons shared by only a subset of instances, i.e. family resemblance. Note that the described example – exclusively fully shared semantic features for concrete and only partially shared ones for abstract concepts – are extreme cases for the purpose of exhibition; in reality, most concepts lie on a continuum between the extremes, being however held together by features with different degrees of sharedness (high for concrete, low for abstract terms).

6.3. Neurocomputational modelling of concrete and abstract concepts and meaning

This structural difference between feature overlap and family resemblance can be used in network simulations to obtain clues about the formation of mechanisms putatively underlying the knowledge about, and processing of, concrete and abstract concepts along with the differences between them. To this end, the brain-constrained semantic model including 12 areas and spiking neurons was used (Tomasello et al., 2018). As mentioned before, each of the 12 areas imitated frontal, temporal, parietal or occipital cortical areas situated in inferior and lateral motor as well as auditory and ventral visual systems, along with their respective connector hub areas (see Fig. 3). Conceptual learning and grounding were simulated by activating neuronal patterns for conceptual instances in primary visual and lateral motor cortex. These were thought to simulate, for example, the shape of and typical hand action performed with a hammer, or the visual shape of a beautiful object and consequent bodily activity. For each category, three patterns and instances were learned, whereby each triplet of patterns implemented the structural features of concrete or abstract concepts, i.e. full feature overlap or family resemblance (Fig. 5). Differences between the resultant neuronal circuits are used to draw careful conclusions on the putative biological basis of differences in abstractness.

The panels in the middle row of Fig. 6 illustrate the results of one typical simulation. For concrete concepts, the neurons specifically activated by the three conceptual instances are shown in blue, green and red (see the small dots within the gray squares showing areas). The semantic overlap neurons equally responsive to all instances, which, together, form the core of the concrete category mechanism, appear in black. Interestingly, these are more common in the central connector hub areas (PF_L and AT) than in the stimulated periphery of the network ($M1_L$, V1). The accumulation of semantic neurons in the center of the

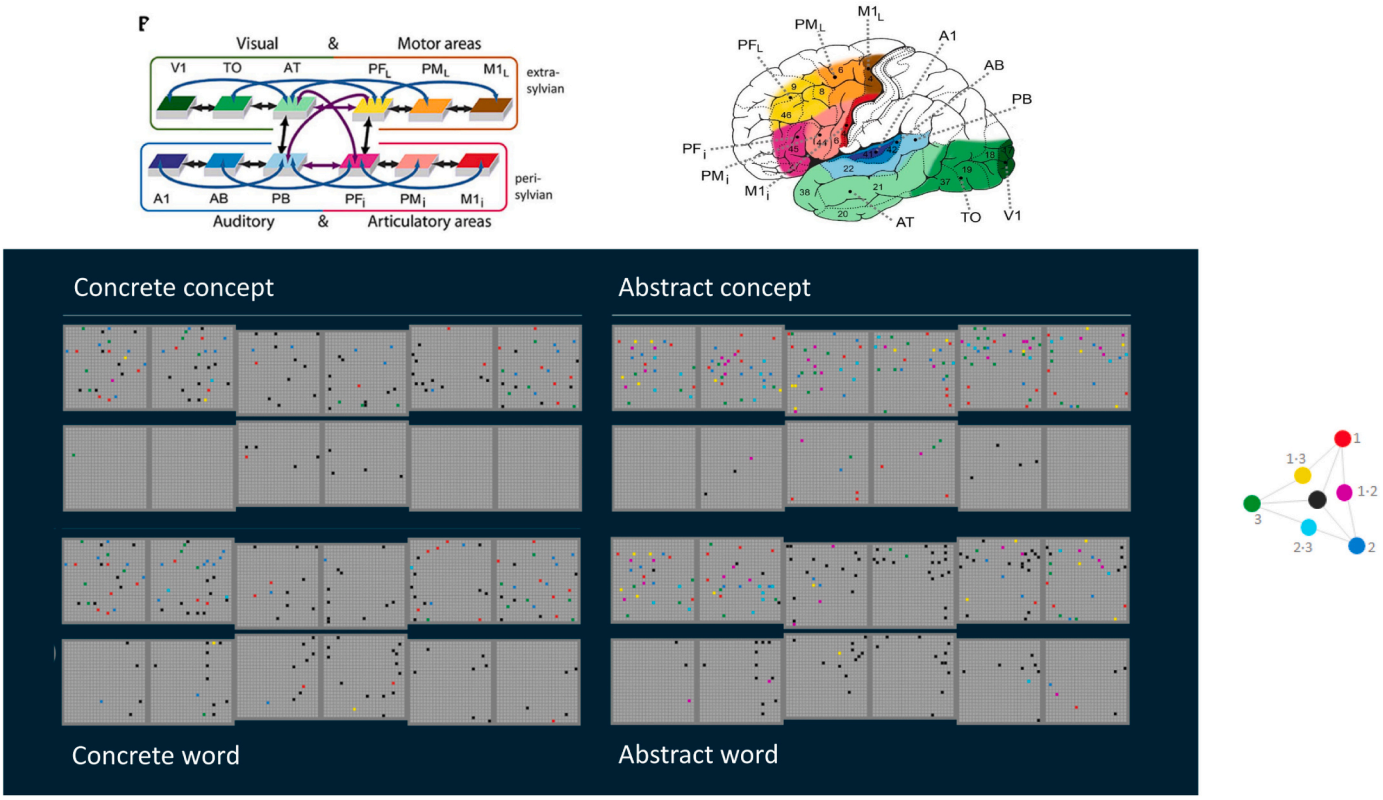


Fig. 6. Modelling of concrete and abstract concept and symbol formation in a brain-constrained model of 12 areas. *Top panels:* The diagram on the left shows the network structure and connectivity as implemented in the 12-area semantic model (cf. Fig. 3). The brain diagram on the right shows the modelled areas in cortex; correspondence is indexed by colors and abbreviations. *Middle panels:* In the conceptual learning experiments, neuronal patterns in visual and motor cortex were activated; these are thought to simulate the perception of and action upon specific objects that fall into a given concrete or abstract conceptual category (e.g., 3 different eyes or 3 different beautiful things), which may contribute to building category representations. Neurons shown in blue, green and red are activated by one grounding instance only, those in cyan, yellow and magenta by 2 and those in black are 'conceptual/semantic neurons' activated by all 3 instances of a concept (see also schematic color diagram on the right). Grounding patterns for concrete concepts – see *middle and bottom panels on the left* – share several features and therefore neurons (as in Fig. 5, panels on the left). Grounding patterns of abstract concepts – see *middle and bottom panels on the right* – show partial feature sharing and family resemblance (as in Fig. 5, panels on the right). As a result of feeding the grounding patterns into the network, distributed circuits develop that correspond to the individual conceptual instances. These overlap for instances of concrete concepts (note the many black neurons, especially in the central areas), whereas for instances of abstract concepts there is partial feature and neuronal overlap (neurons in cyan, yellow, magenta). *Bottom panels:* After learning, grounding patterns of concrete and abstract concepts in the context of spoken words, both concrete and abstract symbol representations, include a majority of neurons (in black) shared across instance representations. (For further discussion, see text and Henningsen-Schomers et al., 2023; Henningsen-Schomers and Pulvermüller, 2022).
(The bottom panels are reproduced with permission by Fynn Dobler from https://www.geisteswissenschaften.fu-berlin.de/en/v/matco/dataviz/label-nolabel/index.html)

architecture (Fig. 6, left middle panel) indicates that the cell assemblies formed during concrete category learning are wellconnected, suggesting that a solid conceptual representation has formed (see also part 4.2 above and Henningsen-Schomers et al., 2023; Henningsen-Schomers and Pulvermüller, 2022). This contrasts with the result of learning abstract concepts based on instances exhibiting family resemblance and thus only partial feature-neuron overlap (Fig. 6, left bottom panel). Neurons responsive to all conceptual instances (shown in white) are rare (which is 'forced upon' the network by the family-resemblance patterns), but those responsive to 2 instances (in cyan, magenta or yellow) are most frequently present in the primary areas. In contrast to the predominance of semantic neurons in the central areas seen for concrete concepts, these central-area semantic neurons (responsive to 2 or all 3 instances) are relatively rare for abstract concepts (Henningsen-Schomers et al., 2023; Henningsen-Schomers and Pulvermüller, 2022), suggesting that a joint activation of these semantic neurons within the network – the network correlate of the processing of an abstract concept – might be difficult to achieve. Indeed, a corresponding difference between the emergent mechanisms for concrete and abstract concepts is seen in activation dynamics (Fig. 7, broken lines). There is substantial activation of the semantic overlap neurons for each of the instances of concrete concepts (peak of the dotted blue curve), whereas the instances of abstract concepts do not substantially activate semantic neurons (low peak of dotted red line, Dobler et al., 2023). In addition, activity is maintained for several time steps by concrete conceptual circuits, but falls off steeply after presentation of abstract conceptual instances (see downslopes of dotted curves). These results suggest that concrete concepts are learned well from experience with their instances, whereas such learning is limited and possibly deficient for abstract ones.

Intriguingly, verbal symbols used concordantly with the instances of a category exert a profound influence on the mechanisms of abstract and concrete concept formation. The bottom panels in Fig. 6 show activation patterns for three instances of concrete and abstract concepts after these instance sets had been associated with their respective category term (Dobler et al., 2023; Henningsen-Schomers et al., 2023). First, it is apparent that many neurons in perisylvian areas (see lower rows of 6 boxes) participate in the activations; this is a straightforward consequence of the link between the categorical instances and the category term. More surprisingly, the neuronal elements shown in white (which, as indicated, index activation to each of the three instances) now seem similarly numerous in concrete and abstract instance processing (Henningsen-Schomers et al., 2023). Consistent with this observation, the

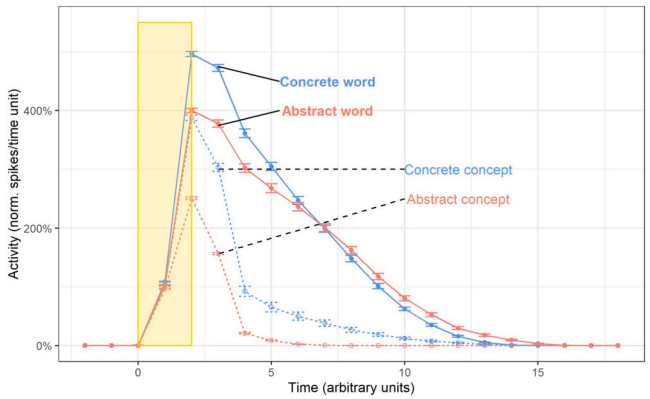


Fig. 7. Activation of shared conceptual/semantic neurons elicited by individual instances of concrete (in blue) and abstract concepts (in red) learned as such (broken lines) and in the context of verbal symbols (solid lines). After conceptual learning out of verbal context, there is significantly stronger activation to concrete than to abstract conceptual instances. In addition, the activation of concrete semantic neurons is maintained for some time (ca. 10 time steps), whereas that for abstract ones decreases immediately after stimulation (after 3 time steps; NB that the path length from one end of the network to the other is 3). Much stronger and comparable activation slopes emerge after concrete and abstract conceptual learning in context of verbal symbols. A high peak activation is followed by an elongated period of activity maintenance. These result show that the processes of cell assembly ignition and reverberation follow concrete conceptual learning outside and in verbal context. For abstract conceptual learning, these processes only emerge in verbal context, suggesting that the building of discrete representations for abstract concepts requires verbal symbolic support. Activity values are normalized to account for differences in the maximal numbers of semantic neurons.
(Adapted from Dobler et al., 2023).

activation dynamics of semantic overlap neurons is now comparable between concept types, showing large activation peaks (Fig. 7, solid lines). The enlarged and similarly pronounced activations for concrete and abstract concepts persist after controlling for the addition of the 'verbal label' (by subtracting the activation curves for separately learned symbol forms from those induced by the meaningful symbols, Dobler et al., 2023).

In essence, the brain-constrained model of concrete semantic category formation shows reasonable learning of concepts based on and induced by the experience of real-world entities, including objects and actions that share perception- and action-related features, which therefore become conceptual or semantic. The learning of a verbal symbol for the category, which is grounded in experiencing the symbol in the context of conceptual instances, builds a higher-order semantic circuit by connecting the conceptual and the symbolic circuits with each other (as schematically illustrated in Fig. 5, bottom left). The resultant semantic circuit is distributed across sensory, motor and language areas and strongly involves the connector hubs interlinking dorsal and ventral frontal and posterior systems. Due to the label addition, there is substantially enhanced activation, which is interpretable as a mechanistic basis of and explanation for the enhancement of attention to referent objects brought about by knowledge of a 'verbal label' for them (Fig. 7, see also Section 5).

In contrast, the brain-constrained implementation of the family-resemblance model of abstract concepts and semantics shows little evidence for abstract concept formation when 'experiencing' real world instances of that concept. The variability and reduced correlation of neuronal unit activations caused by conceptual instances fail to build a strongly connected conceptual cell assembly circuit. As a result, the putative conceptual mechanisms are only manifest in variable and weak activation, so that it remains questionable whether categories have indeed been learned by experience. In contrast, after the same instances have been learned in the context of a category label, there is clear evidence for strong and solid conceptually- and semantically-related activation. With symbol support, a conceptual-semantic circuit has formed, which produces strong, robust and lasting activation. In this context, it is remarkable that, although the family resemblance pattern with only partially shared features across instances was the basis of learning, the finally developing semantic circuit (after label information had been added) included a majority of neuronal elements equally responsive to all instances – not just the family-resemblance subset. This is not explained by adding neural elements for the label, because many of the units responsive to all abstract category instances had been specific to only one or two category members before linguistic learning. Therefore, the network appears to enhance perceptual feature neurons to become semantic, category general elements, although the referent objects and actions that form the basis of the abstract category do, in fact, not all share the related features. This implies a fundamental change of abstract semantic representations related to symbol learning. Thus, the brain-constrained network builds abstract category circuits with semantic neurons that are equally responsive to all category members, although the conceptual instances themselves do not share the related features.

These results sit well with observations from Section 5 about the effects of language on other cognitive domains. The neurobiologically founded model simulations support the hypothesis that symbols change the way humans perceive objects and build concepts (see also Section 7). However, it needs to be kept in mind that the results of these simulations crucially depend on how cognitive learning was simulated and, of course, on the implementation of relevant constraints. Should it turn out that relevant mistakes were made in modeling concept and word learning, that the full/partial feature sharing model is deficient or that relevant biological constraints were omitted, the results and conclusions need to be revised. However, such revision should consist in introducing and applying more realistic constraints on network structure, network functionality and learning and activation procedures.

7. Summary and outlook

The aim of this text is to show that it is necessary – and to illustrate how it is possible – to work towards neurobiological explanations of cognitive, conceptual and symbolic processes, their brain 'loci', activity dynamics and mental characteristics. In this endeavor, neurocomputational modelling using neural networks that are biologically constrained at multiple levels are of special relevance, as only they can deliver detailed insight into the mechanistic correlates underlying aspects of cognition within a brain-like device. The mechanisms discovered in these artificial networks are, of course, not a proof of the existence of analogues mechanisms in the human brain. However, they can provide clues for neurobiologically founded theorizing and for developing biologically plausible mechanistic explanations of cognition. These are not at the level of abstract theorizing, but, instead at the level of 'material' mechanisms, that is, mechanisms that govern the materials involved, i.e. nerve cells, their connections, group-wise interactions and resultant global dynamics.

In this review, the strategy to build and apply brain-constrained models to symbol and concept learning was highlighted. In contrast to most current work with neural networks, this perspective promotes network models implementing a broad range of features that make these models similar to the (human) brain, by implementing neurons, their local connectivity and interaction, biologically realistic learning, regulation mechanisms along with area structure and long-distance connectivity (Section 2). This final section now summarizes main results featured in the review and discusses advantages, limitations and future research needs as well as novel perspectives.

7.1. Are brain-constraints indeed necessary for explaining cognitive mechanisms?

One may reject the main claim that brain-constrained modelling is necessary by stating that an algorithmic explanation of cognitive mechanisms is sufficient. This may appear as particularly plausible if this algorithmic explanation leads to near-perfect imitation of human performance and even correctly predicts brain activation patterns that index specific cognitive activities. In fact, as mentioned in the introduction above, some deep neural networks have recently been shown to generate not only human-like behaviors (e.g., visual object classification or next-word predictions), but, in addition, brain activity patterns consistent with imaging and/or neurophysiological data, although the networks applied here were not particularly well matched to neuroanatomical features of the primate or human brain and used algorithms far from Hebbian or other forms of biologically plausible learning (see, for example, Caucheteux et al., 2021; Caucheteux et al., 2023; Schrimpf et al., 2021; Schrimpf et al., 2020). Instead of biological realism, model optimization aimed at efficient and fast processing to minimize computation time. These models outperformed other types of neural networks, including convolutional and recurrent deep neural networks, which, as discussed above, implement some biological features, including topographic and reverberant connections (see Section 1, Kietzmann et al., 2019b; LeCun et al., 2015). As one prominent example, generative pretrained transformer (GPT) networks lack both features, but, nevertheless, performed best on cognitive-behavioral tasks and in predicting cognitive brain activation patterns (see Schrimpf et al., 2021). And it is obvious that these networks optimized for specific tasks (such as prediction or classification) will also outperform networks fashioned according to features of the human brain. Isn't it, therefore, sufficient to model behavior and brain activity using the most efficient machine learning algorithms, leaving aside any structural and functional internal similarities between learning devices and the neurobiological substrate?

The answer depends on the main aim of the investigation. If predictions on behavior and brain activation are in focus, the most efficient machine learning techniques will naturally win. If the aim is to understand how the brain mechanistically supports and enables cognition, these techniques may or may not be helpful. It is clear that different algorithms can be used to describe and model the same data set. To choose a concrete example from Vaswani et al.s' famous paper, their 'attention' values can be calculated in an additive or multiplicative (vector dot product) manner; the authors chose the latter algorithm, because it is faster and more efficient (Vaswani et al., 2017). Assuming that both algorithm types, additive and multiplicative, lead to comparable results, there would be more than one 'explanatory' algorithm. And one may add, that, algorithmically, even large-number multiplication itself can be realized quite differently, for example by copying from a lookup table, by digit-by-digit multiplication and summing up the results, or by transforming numbers into logarithms, addition and back-transformation. In principle, an unlimited number of algorithms can explain a given set of data points. The 'curse of multiple algorithms' makes it impossible to draw a decision on which - out of a range of alternative algorithms describing a given data set - is ultimately right or wrong insofar as it reflects biological computational reality. To show neurobiological adequacy, an algorithm needs, in addition, to relate to neuronal mechanisms (see, for example, mechanisms capturing Hebbian learning, Artola et al., 1990; Bi and Poo, 2001; Bienenstock et al., 1982). This could help to decide between algorithms and to choose one that fits the brain best. Furthermore, the focus in machine learning on efficient coding and reduction of computation time brings in an important bias: The more efficient the algorithm, the more data can be used to train the algorithm. And, quite obviously, a network that can take in data from the entire world wide web has a better basis for predicting the next words in a text, and for solving other tasks, than one with limited access to a small text corpus. Therefore, model success co-depends on, and may be confounded by, the size of the data base and thus algorithm-efficacy.

These arguments provide further support for a main claim immanent to the research stream summarized in this review: That, in order to find out which neuronal mechanisms underly higher and human cognition, the algorithmic level must be complemented with structural and functional biological constraints (Deco et al., 2011, 2013; Dwivedi et al., 2021; Hahn et al., 2019; Kumar et al., 2010; O'Reilly, 1998; Palm, 1982, 2016; Pulvermüller et al., 2014; Pulvermüller et al., 2021; van Albada et al., 2022; Wennekers et al., 2006). Only if the model does not only produce data exhibiting 'external' similarity to cognitive performance but, in addition, 'internally' resembles brain structure and function and thus uses neuronal circuits with similar structural and functional properties to those that can reasonably be assumed to exist in the (human) brain, only in this case is there hope to overcome the curse of multiple possible algorithms. In Section 2 of this paper, a specific set of constraints was proposed that can make networks structurally and functionally similar to brain networks (see also Pulvermüller et al., 2021). This list may need further specification and extension, but satisfying this or a similar set of constraints appears to be necessary for reaching a mechanistic explanation of the brain's neurocognitive machinery.

7.2. The distributed and discrete nature of conceptual representations

Semantic and conceptual learning has previously been found to lead to fully-distributed and non-discrete patterns for the learned entities. However, recent work suggests that these results may be due to the fact that the networks previously employed lacked realistic local and between area connectivity constraints. A range of studies show that brain-constrained network models implementing local and global connectivity features and realistic Hebbian learning (along with the other constraints of Section 2) build discrete and distributed neuronal cell assemblies for concepts and meaning. They can be said to develop network correlates of symbols and words, although these activate differently, depending on context (Section 3).

Section 4 shows how brain-constrained models can be applied to address additional long standing and cutting-edge questions in the cognitive and linguistic sciences. This was illustrated by addressing specifically human verbal working memory and the related ability to build a huge vocabulary of tens of thousands of symbols (part 4.1). Verbal working memory formation was, in turn, traced back to long-distance inter-area connectivity structure of the human brain (its left-perisylvian cortex) and the resultant temporal dynamics of reverberant cell assembly activity. The question why specific areas in human cortex, in particular in temporal cortex, are generally important for conceptual and semantic processing and why lesions there cause massive semantic impairment was explained based on the distribution of semantic circuits (part 4.2). These circuits, which interlink information about symbol form and meaning, involve neurons in a range of sensory and motor areas in frontal, temporal and occipital lobes. Semantic neuron density was found to be highest in connector hub areas interlinking sensory and motor systems involved in symbol and semantic learning. It was argued that this is because the rich connectivity of these hubs along with their central position in the network leads to accumulating and most persistent neural activation there, resulting in most frequent neuronal co-activation and thus the largest number of neurons being recruited there. Likewise, the question why lesions in modality-preferential sensory and motor areas can lead to subtle category-specific semantic deficits was answered based on differences in the neuron distributions of the emerging semantic circuits across model areas. These topographical differences were caused by the type of correlated information relevant in the semantic grounding process. Furthermore, it was shown that fast semantic mapping between representations of symbolic form and meaning is compatible with and explained by Hebbian learning in a brain-constrained architecture (part 4.3).

Section 4 also showed how, in addition to the 'internal' constraints

by which artificial neural networks are tailored to fit aspects of brain structure and function, behavioral and brain data can be used to evaluate these models 'externally'. This was illustrated using cortical lesion loci and the semantic deficits these typically cause and the generally known (but formerly still unexplained) facts of human-specific verbal working memory and fast semantic mapping. Therefore, behavioral performance patterns, as seen, for example, in neurological disease or after deprivation, are explained by neural models with brain-similarity (e.g., Efremov et al., 2022; Ralph et al., 2017; Stefaniak et al., 2019; Tomasello et al., 2019). Brain activity during object and symbol perception and understanding as recorded in EEG, MEG or fMRI recordings can also be used for external model evaluation (e.g., Garagnani et al., 2017; Khaligh-Razavi and Kriegeskorte, 2014; Kietzmann et al., 2019b; Tomasello et al., 2017). Modelling and explaining behavioral results from developmental studies was in focus in Section 5. The models discussed above have not been tested with large data sets recorded across a range of different tasks (see, for example, Schrimpf et al., 2020), but such testing is clearly an important aim for the future.

7.3. Symbols as a causal factor in cognition and concept formation

In Sections 5 and 6, novel explanations were proposed for the complex interplay between language and attention mechanisms and for the much-debated difference between concrete and abstract concepts and meanings. It is argued that, using biologically constrained mechanistic models, it is possible to explain surprisingly sophisticated cognitive processing, such as the guidance of attention to different sets of object features induced by object-specific proper names and by category terms applicable to an entire class of instances. The Hebbian unsupervised learning algorithm and, in particular, both its associative and its dissociative terms, were shown to underlie and explain the differential binding between proper names and the specific features of their referent objects and between category terms and the semantic features shared across their category instances (Section 5).

Proponents of associative learning may use these results to argue that their approach makes an additional cognitive level of description obsolete and that the principles manifest in learning behavior are sufficient for explaining cognition – or relevant parts of it. However, it needs to be seen that the current approach is markedly different from behaviorist attempts. As explained in detail, associations between words, actions and objects are not sufficient for building the outlined differential mechanisms for proper names and category terms. In order to arrive at the proposed explanations, it is necessary to elaborate on the neuronal microstructure of cognitive entities, their features and shared properties as well as on those properties that make entities relatively unique within a family of similar ones (Pulvermüller, 2018c). Only a detailed neuronal model of cognitive, including conceptual and semantic processes and mechanisms can lead to the exemplified account of proper names and category terms. Such cognitive-neuronal modelling is outside the realm of strictly behavioral approaches.

One may equally well argue that this model just underpins cognitive theories in the structuralist tradition, which long postulated feature-based representations in various domains, the phonological and semantic fields included (Löbner, 2013). These cognitive and linguistic concepts and theories are certainly important. The point to add is that neither the behavioral nor the cognitive-only strategy alone is sufficient. Theories of learning and of cognitive structure are relevant and need to be integrated with each other and with biological principles in order to build explanatory models of symbols, concepts and meaning. The resultant neurobiological mechanisms may be interpreted as the main players determining cognitive function. Alternatively, one may prefer to revert to the metaphor of the child, or cognizing individual more generally, as a 'theorist' (see LaTourrette and Waxman Sandra, 2020; Perszyk and Waxman, 2018; Waxman and Gelman, 2009). After all, any theorist needs a knowledge basis on which theorizing can operate, which, in turn, requires a foundation in brain mechanisms.

Section 6 of the article focused on concreteness vs abstraction/ness. A structural qualitative difference between concrete and abstract concepts was shown to lie in the presence of shared semantic features characterizing all, or at least most, members of concrete conceptual categories and their absence for abstract ones. Instead, the real-world scenes, objects and entities, to which abstract terms apply, differ widely in their perceptual and action-related features and just show family resemblance, that is, partial sharing of features across instances. In brain-constrained models, this difference was found to underlie the findings that concrete categories can be learned from experience, whereas this was not possible for abstract concepts. Support from symbols consistently used with the variable instantiations of an abstract concept was needed for abstract concept building. Furthermore, the brain-constrained model also explained enhanced attention to reference objects after label learning. These results may help to explain why some conceptual categories can be built from experience alone, whereas, for others, this is less likely or even impossible (Thériault et al., 2018).

The results summarized in Sections 5 and 6 also provide a mechanistic biological basis for claims about a role of language in perceptual discrimination and concept formation and, more generally, for addressing causal influences of language on perception and thought. Note that there is meanwhile strong evidence for such causal effects from empirical and experimental studies in the tradition of what is called 'linguistic relativity' (Kemmerer, 2022; Lupyan et al., 2020; Majid et al., 2004; Miller et al., 2018; Thierry, 2016; Thierry et al., 2009; Vanek et al., 2021). The introduced models and brain-constrained network implementations illustrate mechanisms by which language directs attention to perceptual or conceptual features of objects and how symbols assist the building of abstract concepts. These mechanisms were spelled out at the level of the underlying materials, that is, neurons arranged in local clusters and interconnected by local and long-distance connections and merged into distributed functionally discrete circuits. Based on these mechanisms, explanations for causal effects of language on cognition were proposed, which are based on Hebbian association and dissociation learning effective in a multi-area model architecture.

7.4. Limitations, research needs, and perspectives

It is obvious that the results summarized here only cover a small set of cognitive phenomena and their putative explanations using brain-constrained models, with focus on symbol and concept learning and processing. In future, it is desirable to extend this approach to other cognitive domains, including phenomena as different as sentence and construction building, social and communicative interaction, numerical and mathematical skills, or influences of language on object discrimination and action execution. Also, within the symbol processing domain, many questions await addressing, including, for example, the interplay between direct semantic grounding and indirect grounding of novel word forms in the context of co-processed familiar ones.

Apart from a broadening of the topics to be addressed, the level of detail of neurobiologically realistic modelling needs to be increased. For example, the semantic model currently used in several of the reviewed simulations (Figs. 3 and 6) only included model areas imitating a small selection of frontal, temporal and occipital areas of the left hemisphere. An extension of this model to include additional brain structures would be desirable, in particular the right hemisphere, parietal areas and generally a larger set of cortices and subcortical nuclei. At the local microcircuit level, there is further room for increasing brain similarity of the models. Although the local interplay between excitatory and inhibitory neurons is implemented, other models implement such local micro-circuits in much greater detail (see, for example, Schmidt et al., 2018b,a; van Albada et al., 2022). However, as mentioned above, adding detail to the simulations comes with huge additional costs in terms of simulation time and computational resources. Note, for example, that the aforementioned models focusing on detailed local mechanisms include only relatively few local neuron clusters, which is, for example,

far below the 7500 clusters implemented in the twelve areas of the mean-field semantic model (Figs. 3 and 6). Therefore, although model improvement is desirable and necessary, practical limitations need to be equated with the priority degree of each change. Some of these limitations may be mitigated by using or introducing more effective computational techniques or computational resources.

A further limitation of brain-constrained models concerns the main target of this kind of research: To work towards mechanistic explanation of cognitive capacities. If the artificial neural networks are relatively small toy devices including only an overlookable number of neural units, it seems plausible that the mechanisms in these networks are easy to extract, illustrate and use for developing explanations. However, if the number of neurons is increased to thousands or tens of thousands of neurons, specific techniques are needed to extract the functional and structural changes that follow learning. For example, in some of the reviewed examples, such middle size models were used and cell assembly dynamics was extracted to map the formation of network correlates of symbols and meanings. With large-scale models of millions or even billions of neurons, the degree of complexity may become so high that finding mechanisms and correlates of cognitive processes becomes as difficult as in the real brain. In the attempt to model aspects of cognition, it may therefore be advantageous to focus on toy or middle-size models first. At a later step, it will still be desirable to confirm any findings from these with more voluminous and sophisticated models of larger parts of cortex and, ideally, the entire brain.

Extracting from brain-constrained neural networks the putative mechanistic correlates of cognitive processes is only a first step in the endeavor of explaining them neurobiologically. As a second step, it is important to investigate and find out how these putative neuronal correlates dynamically develop over time and interact after they have formed in the different components of the network. Here, the functional features of activation patterns and structural-anatomical differences between areas may provide clues about possible causal factors. In a third step, hypotheses about the causal factors underlying specific phenomena (as, for example, maintenance of activity or particularly high densities of neurons of a specific kind in a particular area) can be formulated. Finally, alternative explanatory hypotheses can be tested against each other by parameter variation in the model. There still remains much to do at this latter level, as several of the aforementioned proposals still allow for more specific hypotheses which can be tested against each other. For example, the areas where working memory and highest circuit neuron densities are observed are characterized by both centrality and high degree of connectivity (Section 4). Whether one or both of these factors are relevant still needs to be investigated. Likewise, the role of Hebbian association and dissociation learning was highlighted in concrete and abstract concept formation (Section 6), but the anatomical makeup or 'depth' of the multi-area network may play an equally important role for the semantic representations that develop.

Still, in spite of the outlined shortcomings, the enterprise of modelling cognition with networks approximating, as well as possible, the neurobiological basis of the cognitive machinery in our brains, seems worth the effort. For sure it offers putative neurobiological explanations for hitherto unexplained cognitive phenomena (e.g., large vocabularies, semantic hub location, attention guidance by labels, abstract conceptual mechanisms), which the classic labelling approach does not address or provide, and a purely algorithmic modelling approach cannot deliver. The strategy of brain-constrained modelling and systematically evaluating the processes and mechanisms that emerge within biologically grounded networks offers novel future perspectives on neurobiologically founded accounts of human mental life.

Funding

This work was supported by the European Research Council through the Advanced Grant "Material constraints enabling human cognition, MatCo" (ERC-2019-ADG 883811), by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany's Excellence Strategy through the Cluster of Excellence "Matters of Activity. Image Space Material, MoA" (DFG EXC 2025/1 - 390648296) and by DFG and Agence Nationale de la Recherche (ARN) research grant "Phonological Networks, PhoNet" (DFG Pu 97/25-1).

Declaration of Competing Interest

The author declares no competing interests.

Data Availability

No novel data was used for the research described in the article. For data availability, please see the cited references.

Acknowledgements

I would like to thank the following colleagues for support and discussion when writing this ms and for comments and suggestions of previous versions of the text: Ad Aertsen, Fynn Dobler, Max Garagnani, Malte Henningsen-Schomers, Johanna Knechtges, Andreas Knoblauch, Phuc Nguyen, Günther Palm, Guillaume Thierry, Rosario Tomasello, Thomas Wennekers, and two anonymous referees.

References

Aleksandrov, A.A., Memetova, K.S., Stankevich, L.N., Knyazeva, V.M., Shtyrov, Y., 2020. Referent's Lexical Frequency Predicts Mismatch Negativity Responses to New Words Following Semantic Training. J. Psycholinguist. Res 49, 187–198.
Allport, D.A., 1985. Distributed systems, modular subsystems and dysphasia. In: Newman, S.K., Epstein, R. (Eds.), Current perspectives in dysphasia. Churchill Livingstone, Edinburgh, pp. 207–244.
Alston, W.P., 1964. Philosophy of language. Prentice-Hall,, Englewood Cliffs, NJ.
Althaus, N., Mareschal, D., 2014. Labels direct infants' attention to commonalities during novel category learning. PLoS One 9, e99670.
Anderson, A.J., Lalor, E.C., Lin, F., Binder, J.R., Fernandino, L., Humphries, C.J., Conant, L.L., Raizada, R.D.S., Grimm, S., Wang, X., 2018. Multiple Regions of a Cortical Network Commonly Encode the Meaning of Words in Multiple Grammatical Positions of Read Sentences. Cereb. Cortex.
Ardesch, D.J., Scholtens, L.H., Li, L., Preuss, T.M., Rilling, J.K., van den Heuvel, M.P., 2019. Evolutionary expansion of connectivity between multimodal association areas in the human brain compared with chimpanzees. Proc. Natl. Acad. Sci. USA 116, 7101–7106.
Artola, A., Singer, W., 1993. Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. Trends Neurosci. 16, 480–487.
Artola, A., Bröcher, S., Singer, W., 1990. Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. Nature 347, 69–72.
Baddeley, A., 2003. Working memory: looking back and looking forward. Nat. Rev. Neurosci. 4, 829–839.
Baddeley, A., Gathercole, S., Papagno, C., 1998. The phonological loop as a language learning device. Psychol. Rev. 105, 158–173.
Baker, G.P., Hacker, P.M.S., 1984. Language, sense and nonsense. Basil Blackwell,, Oxford.
Baker, G.P., Hacker, P.M.S., 2009. Wittgenstein: Understanding and meaning, part 1 - essays. Wiley-Blackwell,, Oxford, Chichester.
Balaban, M.T., Waxman, S.R., 1997. Do words facilitate object categorization in 9-month-old infants? J. Exp. Child Psychol. 64, 3–26.
Baldwin, D.A., Markman, E.M., 1989. Establishing word-object relations: a first step. Child Dev. 60, 381–398.
Barlow, H., 1972. Single units and cognition: a neurone doctrine for perceptual psychology. Perception 1, 371–394.
Barrett, R.L.C., Dawson, M., Dyrby, T.B., Krug, K., Ptito, M., D'Arceuil, H., Croxson, P.L., Johnson, P.J., Howells, H., Forkel, S.J., Dell'Acqua, F., Catani, M., 2020. Differences in Frontal Network Anatomy Across Primate Species. J. Neurosci. 40, 2094–2107.
Barsalou, L.W., 2008. Grounded cognition. Annu Rev. Psychol. 59, 617–645.
Barsalou, L.W., Wiemer-Hastings, K., 2005. Situating abstract concepts. In: Pecher, D., Zwaan, R. (Eds.), Grounding cognition: the role of perception and action in memory, language, and thought. Cambridge University Press, New York, pp. 129–163.
Barsalou, L.W., Dutriaux, L., Scheepers, C., 2018. Moving beyond the distinction between concrete and abstract concepts. Philos. Trans. R. Soc. Lond. B Biol. Sci. 373.
Bates, E., Wilson, S.M., Saygin, A.P., Dick, F., Sereno, M.I., Knight, R.T., Dronkers, N.F., 2003. Voxel-based lesion-symptom mapping. Nat. Neurosci. 6, 448–450.
Bennett, M.R., Hacker, P.M., 2006. Language and cortical function: conceptual developments. Prog. Neurobiol. 80, 20–52.
Bernardi, S., Benna, M.K., Rigotti, M., Munuera, J., Fusi, S., Salzman, C.D., 2020. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. Cell 183, 954–967 e921.

Bertolero, M.A., Yeo, B.T.T., Bassett, D.S., D'Esposito, M., 2018. A mechanistic model of connector hubs, modularity and cognition. Nat. Hum. Behav. 2, 765–777.
Bi, G., Poo, M., 2001. Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev. Neurosci. 24, 139–166.
Bi, G.Q., Poo, M.M., 1998. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J. Neurosci. 18, 10464–10472.
Bibbig, A., Wennekers, T., Palm, G., 1995. A neural network model of the cortico-hippocampal interplay and the representation of contexts. Behav. Brain Res 66, 169–175.
Bienenstock, E.L., Cooper, L.N., Munro, P.W., 1982. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. J. Neurosci. 2, 32–48.
Binder, J.R., Desai, R.H., 2011. The neurobiology of semantic memory. Trends Cogn. Sci. 15, 527–536.
Bion, R.A., Borovsky, A., Fernald, A., 2013. Fast mapping, slow learning: disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30months. Cognition 126, 39–53.
Bishop, D.V., Brown, B.B., Robson, J., 1990. The relationship between phoneme discrimination, speech production, and language comprehension in cerebral-palsied individuals. J. Speech Hear Res 33, 210–219.
Blondin-Massé, A., Harnad, S., Picard, O., St-Louis, B. 2013. Symbol Grounding and the Origin of Language: From Show to Tell. Eds. S. Harnad, Levebre.
Borghi, A.M., Fini, C., Mazzuca, C., 2022a. Abstract Concepts, Social Interaction, and Beliefs. Front Psychol. 13, 919808.
Borghi, A.M., Shaki, S., Fischer, M.H., 2022b. Abstract concepts: external influences, internal constraints, and methodological issues. Psychol. Res.
Borghi, A.M., Shaki, S., Fischer, M.H., 2022c. Concrete constraints on abstract concepts-editorial. Psychol. Res.
Borghi, A.M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., Tummolini, L., 2019. Words as social tools: Language, sociality and inner grounding in abstract concepts. Phys. Life Rev. 29, 120–153.
Borghi, A.M., Mazzuca, C., Da Rold, F., Falcinelli, I., Fini, C., Michalland, A.H., Tummolini, L., 2020. Abstract Words as Social Tools: Which Necessary Evidence. Front Psychol. 11, 613026.
Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332.
Braitenberg, V., 1978. Cell assemblies in the cerebral cortex. In: Heim, R., Palm, G. (Eds.), Theoretical approaches to complex systems. (Lecture notes in biomathematics, vol. 21. Springer, Berlin, pp. 171–188.
Braitenberg, V., Schüz, A., 1998. Cortex: statistics and geometry of neuronal connectivity. Springer,, Berlin.
Braunsdorf, M., Blazquez Freches, G., Roumazeilles, L., Eichert, N., Schurz, M., Uithol, S., Bryant, K.L., Mars, R.B., 2021. Does the temporal cortex make us human? A review of structural and functional diversity of the primate temporal lobe. Neurosci. Biobehav Rev. 131, 400–410.
Brown, R., 1958. Words and things. Free Press,, New York.
Bullmore, E., Sporns, O., 2012. The economy of brain network organization. Nat. Rev. Neurosci. 13, 336–349.
Cangelosi, A., Harnad, S., 2001. The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. Evol. Commun. 4, 117–142.
Cangelosi, A., Stramandinoli, F., 2018. A review of abstract concept learning in embodied agents and robots. Philos. Trans. R. Soc. Lond. B Biol. Sci. 373.
Cangelosi, A., Greco, A., Harnad, S., 2000. From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories1. Connect. Sci. 12, 143–162.
Cangelosi, A., Greco, A., Harnad, S., 2002. Symbol grounding and the symbolic theft hypothesis. In: Cangelosi, A., Parisi, D. (Eds.), Simulating the evolution of language. Springer, London, pp. 3–20.
Caporale, N., Dan, Y., 2008. Spike timing-dependent plasticity: A Hebbian learning rule. Annu Rev. Neurosci. 31, 25–46.
Carey, S., Bartlett, E., 1978. Acquiring a single new word. Papers and Reports on Child Language Development, Number 15, p17–29, Aug 1978 15, 17–29.
Carota, F., Kriegeskorte, N., Nili, H., Pulvermuller, F., 2017. Representational similarity mapping of distributional semantics in left inferior frontal, middle temporal, and motor cortex. Cereb. Cortex 27, 294–309.
Carota, F., Nili, H., Pulvermüller, F., Kriegeskorte, N., 2021. Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: evidence from RSA of BOLD signals. Neuroimage 224, 117408.
Caucheteux, C., Gramfort, A., King, J.-R., 2021. Disentangling syntax and semantics in the brain with deep networks. International conference on machine learning. PMLR,, pp. 1336–1348.
Caucheteux, C., Gramfort, A., King, J.-R., 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat. Hum. Behav. 1–12.
Cazin, N., Llofriu Alonso, M., Scleidorovich Chiodi, P., Pelc, T., Harland, B., Weitzenfeld, A., Fellous, J.M., Dominey, P.F., 2019. Reservoir computing model of prefrontal cortex creates novel combinations of previous navigation sequences from hippocampal place-cell replay with spatial reward propagation. PLoS Comput. Biol. 15, e1006624.
Chen, L., Ralph, M.A.L., Rogers, T.T., 2017. A unified model of human semantic knowledge and its disorders. Nat. Hum. Behav. 1, 0039.
Clahsen, H., 1999. Lexical entries and rules of language: a multidisciplinary study of German inflection. Behav. Brain Sci. 22, 991–1060.
Constant, M., Pulvermüller, F., Tomasello, R., 2023. Brain constrained modelling explains fast mapping of words to meaning. *Cereb Cortex* in press.
Creutzfeldt, O., Ojemann, G., Lettich, E., 1989. Neuronal activity in the human lateral temporal lobe. I. Responses to speech. Exp. Brain Res. 77, 451–475.
de Saussure, F., 1916. Cours de Linguistique Generale. Payot, Paris.
Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech, Lang. Process. 20, 30–42.
Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., Damasio, A.R., 1996. A neural basis for lexical retrieval. Nature 380, 499–505.
Deco, G., Rolls, E.T., 2005. Attention, short-term memory, and action selection: a unifying theory. Prog. Neurobiol. 76, 236–256.
Deco, G., Jirsa, V.K., McIntosh, A.R., 2011. Emerging concepts for the dynamical organization of resting-state activity in the brain. Nat. Rev. Neurosci. 12, 43–56.
Deco, G., Jirsa, V.K., McIntosh, A.R., 2013. Resting brains never rest: computational insights into potential cognitive architectures. Trends Neurosci. 36, 268–274.
Deco, G., Tononi, G., Boly, M., Kringelbach, M.L., 2015. Rethinking segregation and integration: contributions of whole-brain modelling. Nat. Rev. Neurosci. 16, 430–439.
Dell, G.S., 1986. A spreading-activation theory of retrieval in sentence production. Psychol. Rev. 93, 283–321.
Dell, G.S., Schwartz, M.F., Martin, N., Saffran, E.M., Gagnon, D.A., 1997. Lexical access in aphasic and nonaphasic speakers. Psychol. Rev. 104, 801–838.
Dick, A.S., Tremblay, P., 2012. Beyond the arcuate fasciculus: consensus and controversy in the connectional anatomy of language. Brain 135, 3529–3550.
Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., Rekké, S., 2019. Multilink: a computational model for bilingual word recognition and word translation. Biling.: Lang. Cogn. 22, 657–679.
Dobler, F.R., Henningsen-Schomers, M.R., Pulvermüller, F., 2023. Verbal symbols support concrete but enable abstract concept formation: Evidence from brain-constrained deep neural networks. Language Learning, submitted for publication.
Dominey, P.F., Arbib, M.A., 1992. A cortico-subcortical model for generation of spatially accurate sequential saccades. Cereb. Cortex 2, 153–175.
Dominey, P.F., Inui, T., 2009. Cortico-striatal function in sentence comprehension: Insights from neurophysiology and modeling. Cortex 45, 1012–1018.
Doursat, R., Bienenstock, E. 2007. Neocortical self-structuration as a basis for learning. In: Proceedings of the 5th International Conference on Development and Learning (ICDL 2006). pp. 1–6. Indiana University: Bloomington.
Dove, G., 2009. Beyond perceptual symbols: a call for representational pluralism. Cognition 110, 412–431.
Dove, G., 2010. On the need for Embodied and Dis-Embodied Cognition. Front Psychol. 1, 242.
Dove, G., 2016. Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. Psychon. Bull. Rev. 23, 1109–1121.
Dreyer, F.R., Pulvermüller, F., 2018. Abstract semantics in the motor system? - An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. Cortex 100, 52–70.
Dreyer, F.R., Picht, T., Frey, D., Vajkoczy, P., Pulvermüller, F., 2020. The functional relevance of dorsal motor systems for processing tool nouns- evidence from patients with focal lesions. Neuropsychologia 141, 1073–1084.
Drude, L., von Neumann, T., Haeb-Umbach, R. (2018) Deep attractor networks for speaker re-identification and blind source separation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11–15. IEEE.
Dwivedi, K., Bonner, M.F., Cichy, R.M., Roig, G., 2021. Unveiling functions of the visual cortex using task-specific deep neural networks. PLoS Comput. Biol. 17, e1009267.
Efremov, A., Kuptsova, A., Wennekers, T., Shtyrov, Y., Gutkin, B., Garagnani, M., 2022. Simulating semantic dementia in a brain-constrained model of action and object words learning. bioRxiv, 2022.2003. 2003.482066.
Eliasmith, C., Stewart, T.C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., Rasmussen, D., 2012. A large-scale model of the functioning brain. Science 338, 1202–1205.
Elman, J.L., 1990. Finding structure in time. Cogn. Sci. 14, 179–211.
Elman, J.L., 2004. An alternative view of the mental lexicon. Trends Cogn. Sci. 8, 301–306.
Elman, J.L., 2005. Connectionist models of cognitive development: where next? Trends Cogn. Sci. 9, 111–117.
Elman, J.L., Bates, L., Johnson, M., Karmiloff-Smith, A., Parisi, D., Plunkett, K., 1996. Rethinking innateness. A connectionist perspective on development. MIT Press,, Cambridge, MA.
Engel, A.K., Maye, A., Kurthen, M., König, P., 2013. Where's the action? The pragmatic turn in cognitive science. Trends Cogn. Sci. 17, 202–209.
Farah, M.J., McClelland, J.L., 1991. A computational model of semantic memory impairment: modality specificity and emergent category specificity. J. Exp. Psychol.: Gen. 120, 339–357.
Fedorenko, E., Thompson-Schill, S.L., 2014. Reworking the language network. Trends Cogn. Sci. 18, 120–126.
Ferguson, B., Havy, M., Waxman, S.R., 2015. The precision of 12-month-old infants' link between language and categorization predicts vocabulary size at 12 and 18 months. Front Psychol. 6, 1319.
Fischer, M.H., Glenberg, A.M., Moeller, K., Shaki, S., 2021. Grounding (fairly) complex numerical knowledge: an educational example. Psychol. Res.
Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: a critical analysis. Cognition 28, 3–71.
Frege, G., 1892. Über Sinn und Bedeutung. Z. für Philos. und Philos. Krit. 100, 25–50.
Frey, S., Mackey, S., Petrides, M., 2014. Cortico-cortical connections of areas 44 and 45B in the macaque monkey. Brain Lang. 131, 36–55.
Fritz, J., Mishkin, M., Saunders, R.C., 2005. In search of an auditory engram. Proc. Natl. Acad. Sci. USA 102, 9359–9364.

Fulkerson, A.L., Waxman, S.R., 2007. Words (but not tones) facilitate object categorization: evidence from 6- and 12-month-olds. Cognition 105, 218–228.
Fuller, J.A., Burrell, M.H., Yee, A.G., Liyanagama, K., Lipski, J., Wickens, J.R., Hyland, B. I., 2019. Role of homeostatic feedback mechanisms in modulating methylphenidate actions on phasic dopamine signaling in the striatum of awake behaving rats. Prog. Neurobiol. 182, 101681.
Fuster, J.M., 1995. Memory in the cerebral cortex. An empirical approach to neural networks in the human and nonhuman primate. MIT Press,, Cambridge, MA.
Garagnani, M., Pulvermüller, F., 2011. From sounds to words: A neurocomputational model of adaptation, inhibition and memory processes in auditory change detection. Neuroimage 54, 170–181.
Garagnani, M., Pulvermüller, F., 2016. Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. Eur. J. Neurosci. 43, 721–737.
Garagnani, M., Wennekers, T., Pulvermüller, F., 2007. A neuronal model of the language cortex. Neurocomputing 70, 1914–1919.
Garagnani, M., Wennekers, T., Pulvermüller, F., 2008. A neuroanatomically-grounded Hebbian learning model of attention-language interactions in the human brain. Eur. J. Neurosci. 27, 492–513.
Garagnani, M., Wennekers, T., Pulvermüller, F., 2009. Recruitment and consolidation of cell assemblies for words by way of Hebbian learning and competition in a multi-layer neural network. Cogn. Comput. 1, 160–176.
Garagnani, M., Lucchese, G., Tomasello, R., Wennekers, T., Pulvermüller, F., 2016. A spiking neurocomputational model of high-frequency oscillatory brain responses to words and pseudowords. Front Comput. Neurosci. 10, 145.
Garagnani, M., Lucchese, G., Tomasello, R., Wennekers, T., Pulvermüller, F., 2017. A spiking neurocomputational model of high-frequency oscillatory brain responses to words and pseudowords. Front. Comput. Neurosci. 10, 145.
Gebauer, G., 2017. Wie können wir über Emotionen sprechen? In: Gebauer, G., Holodynski, M., Koelsch, S., von Scheve, C. (Eds.), Von der Emotion zur Sprache: Wie wir lernen, über Gefühle zu sprechen. Weilerswist, Velbrück Wissenschaft, pp. 34–84.
Gelman, S.A., Waxman, S.R., 2009. Response to Sloutsky: taking development seriously: theories cannot emerge from associations alone. Trends Cogn. Sci. 13, 332–333.
Gerstner, W., Naud, R., 2009. Neuroscience. How good are neuron models? Science 326, 379–380.
Gerstner, W., Kempter, R., van Hemmen, J.L., Wagner, H., 1996. A neuronal learning rule for sub-millisecond temporal coding. Nature 383, 76–81.
Gibson, J., J., 1979. The Ecological Approach to Visual Perception. Houghton-Mifflin, Boston.
Glasser, M.F., Rilling, J.K., 2008. DTI tractography of the human brain's language pathways. Cereb. Cortex 18, 2471–2482.
Gleitman, L.R., 1990. The structural sources of verb meanings. Lang. Acquis. 1, 3–55.
Glenberg, A.M., 2021. Embodiment and learning of abstract concepts (such as algebraic topology and regression to the mean). Psychol. Res.
Glenberg, A.M., Robertson, D.A., 2000. Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. J. Mem. Lang. 43, 379–401.
Glenberg, A.M., Gallese, V., 2012. Action-based language: a theory of language acquisition, comprehension, and production. Cortex 48, 905–922.
Gliga, T., Volein, A., Csibra, G., 2010. Verbal labels modulate perceptual object processing in 1-year-old children. J. Cogn. Neurosci. 22, 2781–2789.
Grainger, J., Jacobs, A.M., 1996. Orthographic processing in visual word recognition: a multiple read-out model. Psychol. Rev. 103, 518–565.
Graves, A., Mohamed, A.-R., Hinton, G. , 2013 Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE.
Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Wedeen, V.J., Sporns, O., 2008. Mapping the structural core of human cerebral cortex. PLoS Biol. 6, e159.
Hahn, G., Ponce-Alvarez, A., Deco, G., Aertsen, A., Kumar, A., 2019. Portraits of communication in neuronal networks. Nat. Rev. Neurosci. 20, 117–127.
Hale, S.C., 1988. Spacetime and the contrete/abstract distinction. Philos. Stud. 53, 85–102.
Harnad, S., 1990. The symbol grounding problem. Phys. D. 42, 335–346.
Harpaintner, M., Sim, E.J., Trumpp, N.M., Ulrich, M., Kiefer, M., 2020. The grounding of abstract concepts in the motor and visual system: An fMRI study. Cortex 124, 1–22.
Hebb, D.O., 1949. The organization of behavior. A Neuropsychological Theory. John Wiley, New York.
Henningsen-Schomers, M.R., Pulvermüller, F., 2022. Modelling concrete and abstract concepts using brain-constrained deep neural networks. Psychol. Res 86, 2533–2559.
Henningsen-Schomers, M.R., Garagnani, M., Pulvermüller, F., 2023. Influence of language on perception and concept formation in a brain-constrained deep neural network model. Philos. Trans. R. Soc. Lond. B Biol. Sci. 378, 20210373.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.
Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., Botvinick, M., 2021. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. Nat. Commun. 12, 6456.
Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A., 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. International conference on learning representations.
Hodges, J.R., Patterson, K., 2007. Semantic dementia: a unique clinicopathological syndrome. Lancet Neurol. 6, 1004–1014.
Holodynski, M., 2017. Wie Kinder lernen, über ihre Emotionen zu sprechen. In: Gebauer, G., Holodynski, M., Koelsch, S., von Scheve, C. (Eds.), Von der Emotion zur Sprache: Wie wir lernen, über Gefühle zu sprechen. Velbrück Wissenschaft, Weilerswist, pp. 85–189.
Hubel, D., 1995. Eye, brain, and vision. Scientific American Library, New York.
Hubel, D.H., Wiesel, T.N., 1977. Functional architecture of macaque monkey visual cortex (Ferrier Lecture). Proc. R. Soc. Lond.,B 198, 1–59.
Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.
Huyck, C.R., Passmore, P.J., 2013. A review of cell assemblies. Biol. Cyber 107, 263–288.
Ito, T., Klinger, T., Schultz, D., Murray, J., Cole, M., Rigotti, M., 2022. Compositional generalization through abstract representations in human and artificial neural networks. Adv. Neural Inf. Process. Syst. 35, 32225–32239.
Ivanova, M.V., Dragoy, O., Kuptsova, S.V., Yu Akinina, S., Petrushevskii, A.G., Fedina, O. N., Turken, A., Shklovsky, V.M., Dronkers, N.F., 2018. Neural mechanisms of two different verbal working memory tasks: A VLSM study. Neuropsychologia 115, 25–41.
Jackendoff, R., 2002. Foundations of language: Brain, meaning, grammar, evolution. Oxford University Press,, Oxford, UK.
Jackson, R.L., Rogers, T.T., Lambon Ralph, M.A., 2021. Reverse-engineering the cortical architecture for controlled semantic cognition. Nat. Hum. Behav.
Johnston, W.J., Fusi, S., 2023. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. Nat. Commun. 14, 1040.
Katz, J.J., Fodor, J.A., 1963. The structure of a semantic theory. Language 170–210.
Kemmerer, D., 2014. Cognitive Neuroscience of Language. Psychology Press, New York.
Kemmerer, D., 2022. Grounded cognition entails linguistic relativity: a neglected implication of a major semantic theory. Top. Cogn. Sci.
Kemmerer, D., Weber-Fox, C., Price, K., Zdanczyk, C., Way, H., 2007. Big brown dog or brown big dog? an electrophysiological study of semantic constraints on prenominal adjective order. Brain Lang. 100, 238–256.
Kempter, R., Gerstner, W., Van Hemmen, J.L., 1999. Hebbian learning and spiking neurons. Phys. Rev. E 59, 4498.
Keysers, C., Gazzola, V., 2014. Hebbian learning and predictive mirror neurons for actions, sensations and emotions. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369, 20130175.
Khaligh-Razavi, S.M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915.
Kiefer, M., Pulvermüller, F., 2012. Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. Cortex 48, 805–825.
Kietzmann, T., McClure, P., Kriegeskorte, N., 2019a. Deep neural networks in computational neuroscience. Oxford Research Encyclopedia, Neuroscience. Oxford University Press,, Oxford.
Kietzmann, T.C., Spoerer, C.J., Sorensen, L.K.A., Cichy, R.M., Hauk, O., Kriegeskorte, N., 2019b. Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. USA 116, 21854–21863.
Kousta, S.T., Vigliocco, G., Vinson, D.P., Andrews, M., Del Campo, E., 2011. The representation of abstract words: why emotion matters. J. Exp. Psychol. Gen. 140, 14–34.
Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17, 401–412.
Kriegeskorte, N., Diedrichsen, J., 2019. Peeling the onion of brain representations. Annu Rev. Neurosci. 42, 407–432.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 1097–1105.
Kuhnke, P., Kiefer, M., Hartwigsen, G., 2020. Task-dependent recruitment of modality-specific and multimodal regions during conceptual processing. Cereb. Cortex 30, 3938–3959.
Kumar, A., Rotter, S., Aertsen, A., 2010. Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding. Nat. Rev. Neurosci. 11, 615–627.
Lakoff, G., 1987. Women, fire, and dangerous things. What categories reveal about the mind. University of Chicago Press, Chicago.
Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104, 211–240.
Langacker, R.W., 1991. Foundations of Cognitive Grammar. 2 vols. Stanford University Press, Stanford.
Langland-Hassan, P., Faries, F.R., Gatyas, M., Dietz, A., Richardson, M.J., 2021. Assessing abstract thought and its relation to language with a new nonverbal paradigm: Evidence from aphasia. Cognition 211, 104622.
LaTourrette, A.S., Waxman Sandra, R., 2020. Naming guides how 12-month-old infants encode and remember objects. Proc. Natl. Acad. Sci. 117, 21230–21234.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lenci, A., Lebani, G.E., Passaro, L.C., 2018. The emotions of abstract words: a distributional semantic analysis. Top. Cogn. Sci. 10, 550–572.
Levelt, W.J.M., 1989. Speaking. From intention to articulation. MIT Press, Cambridge, MA.
Li, P., Farkas, I., MacWhinney, B., 2004. Early lexical development in a self-organizing neural network. Neural Netw. 17, 1345–1362.
Li, P., Zhao, X., Mac Whinney, B., 2007. Dynamic self-organization and early lexical development in children. Cogn. Sci. 31, 581–612.
Lidz, J., Gleitman, L.R., 2004. Argument structure and the child's contribution to language learning. Trends Cogn. Sci. 8, 157–161.

Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., Hinton, G., 2020. Backpropagation and the brain. Nat. Rev. Neurosci. 21, 335–346.
Lindsay, G.W., Rigotti, M., Warden, M.R., Miller, E.K., Fusi, S., 2017. Hebbian learning in a random network captures selectivity properties of the prefrontal cortex. J. Neurosci. 37, 11021–11036.
Linzen, T., Baroni, M., 2021. Syntactic structure from deep learning. Annu. Rev. Linguist. 7, 195–212.
Liu, N., Kriegeskorte, N., Mur, M., Hadj-Bouziane, F., Luh, W.M., Tootell, R.B., Ungerleider, L.G., 2013. Intrinsic structure of visual exemplar and category representations in macaque brain. J. Neurosci. 33, 11346–11360.
Löbner, S., 2013. Understanding semantics. Routledge,, Oxford.
Locke, J., 1909/1847. An essay concerning human understanding, or, the conduct of the understanding. Kay and Troutman,, Philadelphia.
Löhr, G., 2022. What are abstract concepts? On lexical ambiguity and concreteness ratings. Rev. Philos. Psychol. 13, 549–566.
Lopez-Barroso, D., Catani, M., Ripolles, P., Dell'Acqua, F., Rodriguez-Fornells, A., de Diego-Balaguer, R., 2013. Word learning is mediated by the left arcuate fasciculus. Proc. Natl. Acad. Sci. USA 110, 13168–13173.
Lupyan, G., 2012a. Linguistically modulated perception and cognition: the label-feedback hypothesis. Front. Psychol. 3.
Lupyan, G., 2012b. What do words do? Toward a theory of language-augmented thought. In: Ross, B.H. (Ed.), The psychology of learning and motivation. Elsevier Inc - Academic Press, New York.
Lupyan, G., Mirman, D., 2013. Linking language and categorization: evidence from aphasia. Cortex 49, 1187–1194.
Lupyan, G., Abdel Rahman, R., Boroditsky, L., Clark, A., 2020. Effects of language on visual perception. Trends Cogn. Sci. 24, 930–944.
Machery, E., 2016. The amodal brain and the offloading hypothesis. Psychon. Bull. Rev. 23, 1090–1095.
MacKay, D.G., 1987. The organization of perception and action. A theory of language and other cognitive skills. Springer-Verlag,, New York.
MacNamara, J., 1972. Cognitive basis of language learning in infants. Psychol. Rev. 79, 1–13.
Mahon, B.Z., Caramazza, A., 2008. A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. J. Physiol. Paris 102, 59–70.
Majid, A., Bowerman, M., Kita, S., Haun, D.B., Levinson, S.C., 2004. Can language restructure cognition? the case for space. Trends Cogn. Sci. 8, 108–114.
Majid, A., Roberts, S.G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., Woll, B., LeLan, B., de Sousa, H., Cansler, B.L., Shayan, S., de Vos, C., Senft, G., Enfield, N.J., Razak, R.A., Fedden, S., Tufvesson, S., Dingemanse, M., Ozturk, O., Brown, P., Hill, C., Le Guen, O., Hirtzel, V., van Gijn, R., Sicoli, M.A., Levinson, S.C., 2018. Differential coding of perception in the world's languages. Proc. Natl. Acad. Sci. USA 115, 11369–11376.
Marcus, G., 2018. Deep learning: A critical appraisal. arXiv, 1801, 00631.
Marcus, G.F., 2008. Kluge: The Haphazard Construction of the Human Mind. Houghton Mifflin Co, Boston, MA.
Martin, A., 2007. The representation of object concepts in the brain. Annu Rev. Psychol. 58, 25–45.
Martin, A., 2016. GRAPES—grounding representations in action, perception, and emotion systems: how object properties and categories are represented in the human brain. Psychon. Bull. Rev. 23, 979–990.
Mayor, J., Plunkett, K., 2010. A neurocomputational account of taxonomic responding and fast mapping in early word learning. Psychol. Rev. 117, 1–31.
McClelland, J.L., Rumelhart, D.E., 1985. Distributed memory and the representation of general and specific information. J. Exp. Psychol.: Gen. 114, 159–188.
McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cogn. Psychol. 18, 1–86.
McClelland, J.L., Patterson, K., 2002. Rules or connections in past-tense inflections: what does the evidence rule out? Trends Cogn. Sci. 6, 465–472.
McClelland, J.L., Botvinick, M.M., Noelle, D.C., Plaut, D.C., Rogers, T.T., Seidenberg, M. S., Smith, L.B., 2010. Letting structure emerge: connectionist and dynamical systems approaches to cognition. Trends Cogn. Sci. 14, 348–356.
Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010.
Mesulam, M.M., 2013. Primary progressive aphasia and the language network: the 2013H. Houston Merritt Lecture. Neurology 81, 456–462.
Miller, T.M., Schmidt, T.T., Blankenburg, F., Pulvermüller, F., 2018. Verbal labels facilitate tactile perception. Cognition 171, 172–179.
Morton, J., 1969. The interaction of information in word recognition. Psychol. Rev. 76, 165–178.
Moseley, R., Carota, F., Hauk, O., Mohr, B., Pulvermüller, F., 2012. A role for the motor system in binding abstract emotional meaning. Cereb. Cortex 22, 1634–1647.
Moseley, R.L., Pulvermüller, F., 2018. What can autism teach us about the role of sensorimotor systems in higher cognition? New clues from studies on language, action semantics, and abstract emotional concept processing. Cortex 100, 149–190.
Naumann, D., Frassinelli, D., Schulte im Walde, S. , 2018 Quantitative semantic variation in the contexts of concrete and abstract words. In: Seventh Joint Conference on Lexical and Computational Semantics (SEM 2018). pp. 76–85. Association for Computational Linguistics.
Nguyen, P.T.U., Henningsen-Schomers, M.R., Pulvermüller, F., 2023. Causal influence of linguistic learning on perceptual and conceptual processing: A brain-constrained deep neural network study of proper names and category terms. *J Neurosci, submitted for publication.*
O'Reilly, R.C., 1998. Six principles for biologically based computational models of cortical cognition. Trends Cogn. Sci. 2, 455–562.
Paivio, A., 1971. Imagery and Verbal Processes. Holt, Rinehart and Winston, New York.
Paivio, A., 1991. Dual coding theory: retrospect and current status. Can. J. Psychol. 45, 255–287.
Paivio, A., 2013. Dual coding theory, word abstractness, and emotion: a critical review of Kousta et al. (2011). J. Exp. Psychol. Gen. 142, 282–287.
Palm, G., 1982. Neural Assemblies. Springer, Berlin.
Palm, G., 2016. Neural information processing in cognition: we start to understand the orchestra, but where is the conductor? Front Comput. Neurosci. 10, 3.
Papadimitriou, C.H., Vempala, S.S., Mitropolsky, D., Collins, M., Maass, W., 2020. Brain computation by assemblies of neurons. Proc. Natl. Acad. Sci. USA 117, 14464–14472.
Patterson, K., Nestor, P.J., Rogers, T.T., 2007. Where do you know what you know? the representation of semantic knowledge in the human brain. Nat. Rev. Neurosci. 8, 976–987.
Perszyk, D.R., Waxman, S.R., 2018. Linking language and cognition in infancy. Annu Rev. Psychol. 69, 231–250.
Petrides, M., Pandya, D.N., 2009. Distinct parietal and temporal pathways to the homologues of Broca's area in the monkey. PLoS Biol. 7, e1000170.
Petrides, M., Tomaiuolo, F., Yeterian, E.H., Pandya, D.N., 2012. The prefrontal cortex: comparative architectonic organization in the human and the macaque monkey brains. Cortex 48, 46–57.
Pinker, S., 1994. The Language Instinct. How the Mind Creates Language. Harper Collins Publishers, New York.
Pinker, S., Ullman, M.T., 2002. The past and future of the past tense. Trends Cogn. Sci. 6, 456–463.
Plaut, D.C., 1995. Double dissociation without modularity: evidence from connectionist neuropsychology. J. Clin. Exp. Neuropsychol. 17, 291–321.
Plaut, D.C., Patterson, K., 2010. Beyond functional architecture in cognitive neuropsychology: a reply to Coltheart. Top. Cogn. Sci. 2, 12–14.
Preissler, M.A., Carey, S., 2005. The role of inferences about referential intent in word learning: evidence from autism. Cognition 97, B13–B23.
Pulvermüller, F., 1999. Words in the brain's language. Behav. Brain Sci. 22, 253–336.
Pulvermüller, F., 2005. Brain mechanisms linking language and action. Nat. Rev. Neurosci. 6, 576–582.
Pulvermüller, F., 2013. How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. Trends Cogn. Sci. 17, 458–470.
Pulvermüller, F., 2018a. The case of CAUSE: neurobiological mechanisms for grounding an abstract concept. Philos. Trans. R. Soc. Lond. B Biol. Sci. 373.
Pulvermüller, F., 2018b. Neural reuse of action perception circuits for language, concepts and communication. Prog. Neurobiol. 160, 1–44.
Pulvermüller, F., 2018c. Neurobiological mechanisms for semantic feature extraction and conceptual flexibility. Top. Cogn. Sci. 10, 590–620.
Pulvermüller, F., Fadiga, L., 2010. Active perception: sensorimotor circuits as a cortical basis for language. Nat. Rev. Neurosci. 11, 351–360.
Pulvermüller, F., Garagnani, M., 2014. From sensorimotor learning to memory cells in prefrontal and temporal association cortex: a neurocomputational study of disembodiment. Cortex 57, 1–21.
Pulvermüller, F., Garagnani, M., Wennekers, T., 2014. Thinking in circuits: towards neurobiological explanation in cognitive neuroscience. Biol. Cyber 108, 573–593.
Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M.R., Wennekers, T., 2021. Biological constraints on neural network models of cognitive function. Nat. Rev. Neurosci. 22, 488–502.
Pusch, R., Clark, W., Rose, J., Gunturkun, O., 2023. Visual categories and concepts in the avian brain. Anim. Cogn. 26, 153–173.
Quiroga, R.Q., Kreiman, G., Koch, C., Fried, I., 2008. Sparse but not 'grandmother-cell' coding in the medial temporal lobe. Trends Cogn. Sci. 12, 87–91.
Ralph, M.A., Jefferies, E., Patterson, K., Rogers, T.T., 2017. The neural and computational bases of semantic cognition. Nat. Rev. Neurosci. 18, 42–55.
Rauschecker, J.P., 1991. Mechanisms of visual plasticity: Hebb synapses, NMDA receptors, and beyond. Physiol. Rev. 71, 587–615.
Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., Gillon, C.J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G.W., Miller, K.D., Naud, R., Pack, C.C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A.C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., Kording, K.P., 2019. A deep learning framework for neuroscience. Nat. Neurosci. 22, 1761–1770.
Rilling, J.K., 2014. Comparative primate neuroimaging: insights into human brain evolution. Trends Cogn. Sci. 18, 46–55.
Rilling, J.K., van den Heuvel, M.P., 2018. Comparative primate connectomics. Brain Behav. Evol. 91, 170–179.
Rilling, J.K., Glasser, M.F., Jbabdi, S., Andersson, J., Preuss, T.M., 2011. Continuity, divergence, and the evolution of brain language pathways. Front Evol. Neurosci. 3, 11.
Rilling, J.K., Glasser, M.F., Preuss, T.M., Ma, X., Zhao, T., Hu, X., Behrens, T.E., 2008. The evolution of the arcuate fasciculus revealed with comparative DTI. Nat. Neurosci. 11, 426–428.
Rogers, T.T., McClelland, J.L., 2004. Semantic Cognition. A Parallel Distributed Processing Approach. MIT Press, Cambridge, MA.
Rojas, R., 2013. Neural Networks: A Systematic Introduction. Springer Science & Business Media, Berlin.
Rosch, E., Mervis, C.B., 1975. Family resemblances: studies in the internal structure of categories. Cogn. Psychol. 7, 573–605.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Schmidt, M., Bakker, R., Hilgetag, C.C., Diesmann, M., van Albada, S.J., 2018. Multi-scale account of the network structure of macaque visual cortex. Brain Struct Funct 223, 1409–1435.
Schmidt, M., Bakker, R., Shen, K., Bezgin, G., Diesmann, M., van Albada, S.J., 2018. A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas. PLoS Comput. Biol. 14, e1006359.
Schomers, M.R., Garagnani, M., Pulvermüller, F., 2017. Neurocomputational consequences of evolutionary connectivity changes in perisylvian language cortex. J. Neurosci. 37, 3045–3055.
Schrimpf, M., Kubilius, J., Lee, M.J., Murty, N.A.R., Ajemian, R., DiCarlo, J.J., 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. Neuron 108, 413–423.
Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E., 2021. The neural architecture of language: Integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. 118, e2105646118.
Schwanenflugel, P., Harnishfeger, K.K., Stowe, R.W., 1988. Context availability and lexical decision for abstract and concrete words. J. Mem. Lang. 27, 499–520.
Schyns, P.G., 1991. A modular neural network of concept acquisition. Cogn. Sci. 13, 461–508.
Scott, B.H., Mishkin, M., Yin, P., 2012. Monkeys have a limited form of short-term memory in audition. Proc. Natl. Acad. Sci. USA 109, 12237–12241.
Scott, B.H., Mishkin, M., Yin, P., 2014. Neural correlates of auditory short-term memory in rostral superior temporal cortex. Curr. Biol. 24, 2767–2775.
Searle, J.R., 1980. Minds, brains, and programs. Behav. Brain Sci. 3, 417–457.
Searle, J.R., 1984. Minds, Brains, and Science. Harvard University Press, Cambridge, MA.
Sepulcre, J., Sabuncu, M.R., Yeo, T.B., Liu, H., Johnson, K.A., 2012. Stepwise connectivity of the modal cortex reveals the multimodal organization of the human brain. J. Neurosci. 32, 10649–10661.
Shallice, T., 1988. From Neuropsychology to Mental Structure. Cambridge University Press, New York.
Shebani, Z., Nestor, P.J., Pulvermüller, F., 2021. What's "up"? impaired spatial preposition processing in posterior cortical atrophy. Front Hum. Neurosci. 15, 731104.
Shebani, Z., Patterson, K., Nestor, P.J., Diaz-de-Grenu, L.Z., Dawson, K., Pulvermüller, F., 2017. Semantic word category processing in semantic dementia and posterior cortical atrophy. Cortex 93, 92–106.
Shtyrov, Y., 2011. Fast mapping of novel word forms traced neurophysiologically. Front Psychol. 2, 340.
Sloutsky, V.M., 2009. Theories about 'theories': where is the explanation? comment on waxman and gelman. Trends Cogn. Sci. 13, 331–332.
Sloutsky, V.M., 2010. From perceptual categories to concepts: what develops? Cogn. Sci. 34, 1244–1286.
Sloutsky, V.M., Robinson, C.W., 2008. The role of words and sounds in infants' visual processing: from overshadowing to attentional tuning. Cogn. Sci. 32, 342–365.
Sloutsky, V.M., Yim, H., Yao, X., Dennis, S., 2017. An associative account of the development of word learning. Cogn. Psychol. 97, 1–30.
Smit, P., Virpioja, S., Kurimo, M., 2021. Advances in subword-based HMM-DNN speech recognition across languages. Comput. Speech Lang. 66, 101–158.
Stefaniak, J.D., Halai, A.D., Ralph, M.A.L., 2019. The neural and neurocomputational bases of recovery from post-stroke aphasia. Nat. Rev. Neurol. 16, 43–55.
Steinschneider, M., Fishman, Y.I., Arezzo, J.C., 2003. Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey. J. Acoust. Soc. Am. 114, 307–321.
Stramandinoli, F., Marocco, D., Cangelosi, A., 2017. Making sense of words: a robotic model for language abstraction. Auton. Robots 41, 367–383.
Thériault, C., Pérez-Gay, F., Rivas, D., Harnad, S., 2018. Learning-induced categorical perception in a neural network model. arXiv arXiv:1805.04567.
Thiebaut de Schotten, M., Dell'Acqua, F., Valabregue, R., Catani, M., 2012. Monkey to human comparative anatomy of the frontal lobe association tracts. Cortex 48, 82–96.
Thierry, G., 2016. Neurolinguistic relativity: how language flexes human perception and cognition. Lang. Learn 66, 690–713.
Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., Kuipers, J.R., 2009. Unconscious effects of language-specific terminology on preattentive color perception. Proc. Natl. Acad. Sci. USA 106, 4567–4570.
Tomasello, R., Garagnani, M., Wennekers, T., Pulvermüller, F., 2017. Brain connections of words, perceptions and actions: a neurobiological model of spatio-temporal semantic activation in the human cortex. Neuropsychologia 98, 111–129.
Tomasello, R., Garagnani, M., Wennekers, T., Pulvermüller, F., 2018. A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. Front Comput. Neurosci. 12, 88.
Tomasello, R., Wennekers, T., Garagnani, M., Pulvermüller, F., 2019. Visual cortex recruitment during language processing in blind individuals is explained by Hebbian learning. Sci. Rep. 9, 3579.
Tranel, D., Kemmerer, D., 2004. Neuroanatomical correlates of spatial prepositions. Cogn. Neuropsychol. 21, 719–749.
Tremblay, P., Dick, A.S., 2016. Broca and Wernicke are dead, or moving past the classic model of language neurobiology. Brain Lang. 162, 60–71.
Tsumoto, T., 1992. Long-term potentiation and long-term depression in the neocortex. Prog. Neurobiol. 39, 209–228.
Ueno, T., Saito, S., Rogers, T.T., Lambon Ralph, M.A., 2011. Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. Neuron 72, 385–396.
van Albada, S.J., Morales-Gregorio, A., Bakker, R., Palm, G., Goulas, A., Bludau, S., Dickscheid, T., Hilgetag, C.-C., Diesmann, M., 2020. Bringing anatomical information into neuronal network models. arXiv arXiv:1312.6026.
van Albada, S.J., Morales-Gregorio, A., Dickscheid, T., Goulas, A., Bakker, R., Bludau, S., Palm, G., Hilgetag, C.C., Diesmann, M., 2022. Bringing Anatomical Information into Neuronal Network Models. Adv. Exp. Med Biol. 1359, 201–234.
van den Heuvel, M.P., Sporns, O., 2013. Network hubs in the human brain. Trends Cogn. Sci. 17, 683–696.
Vanek, N., Sóskuthy, M., Majid, A., 2021. Consistent verbal labels promote odor category learning. Cognition 206, 104485.
Vannuscorps, G., Caramazza, A., 2016. Typical action perception and interpretation without motor simulation. Proc. Natl. Acad. Sci. USA 113, 86–91.
Varela, F.J., Thompson, E., Rosch, E., 1991. The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Boston, MA.
Vasilyeva, M.J., Knyazeva, V.M., Aleksandrov, A.A., Shtyrov, Y., 2019. Neurophysiological correlates of fast mapping of novel words in the adult brain. Front Hum. Neurosci. 13, 304.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Verduzco-Flores, S., Bodner, M., Ermentrout, B., Fuster, J.M., Zhou, Y., 2009. Working memory cells' behavior may be explained by cross-regional networks with synaptic facilitation. PLoS One 4, e6399.
Vigliocco, G., Kousta, S.T., Della Rosa, P.A., Vinson, D.P., Tettamanti, M., Devlin, J.T., Cappa, S.F., 2014. The neural representation of abstract words: the role of emotion. Cereb. Cortex 24, 1767–1777.
Vincent-Lamarre, P., Masse, A.B., Lopes, M., Lord, M., Marcotte, O., Harnad, S., 2016. The latent structure of dictionaries. Top. Cogn. Sci. 8, 625–659.
Vohryzek, J., Cabral, J., Castaldo, F., Sanz-Perl, Y., Lord, L.D., Fernandes, H.M., Litvak, V., Kringelbach, M.L., Deco, G., 2023. Dynamic sensitivity analysis: Defining personalised strategies to drive brain state transitions via whole brain modelling. Comput. Struct. Biotechnol. J. 21, 335–345.
Warrington, E.K., Shallice, T., 1984. Category specific semantic impairments. Brain 107, 829–854.
Waxman, S.R., Markow, D.B., 1995. Words as invitations to form categories: evidence from 12- to 13-month-old infants. Cogn. Psychol. 29, 257–302.
Waxman, S.R., Braun, I., 2005. Consistent (but not variable) names as invitations to form object categories: New evidence from 12-month-old infants. Cognition 95, B59–B68.
Waxman, S.R., Gelman, S.A., 2009. Early word-learning entails reference, not merely associations. Trends Cogn. Sci. 13, 258–263.
Wennekers, T., Garagnani, M., Pulvermüller, F., 2006. Language models based on Hebbian cell assemblies. J. Physiol. Paris 100, 16–30.
Westermann, G., Mareschal, D., 2014. From perceptual to language-mediated categorization. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369, 20120391.
Westermann, G., Sirois, S., Shultz, T.R., Mareschal, D., 2006. Modeling developmental cognitive neuroscience. Trends Cogn. Sci. 10, 227–232.
Wittgenstein, L., 1953. Philosophical Investigations. Blackwell Publishers, Oxford.
Wood, C.C., 1978. Variation on a theme of Lashley: Lesion experiments on the neural model of Anderson, Silverstein, Ritz & Jones. Psychol. Rev. 85, 582–591.
Wood, C.C., 1980. Interpretation of real and simulated lesion experiments. Psychol. Rev. 87, 474–476.
Yi, H.G., Leonard, M.K., Chang, E.F., 2019. The encoding of speech sounds in the superior temporal gyrus. Neuron 102, 1096–1110.
Yuille, A.L., Geiger, D., 2003. Winner-take-all networks. In: Arbib, M.A., Bradford, A. (Eds.), The handbook of brain theory and neural networks. Book/MIT Press, Boston, MA, pp. 1228–1231.
Zhou, H.-Y., Liu, A.-A., Nie, W.-Z., Nie, J., 2019. Multi-view saliency guided deep neural network for 3-D object retrieval and classification. IEEE Trans. Multimed. 22, 1496–1506.