Dynamical systems implementation of intrinsic lexical meaning

Abstract

This paper proposes a model for implementation of intrinsic lexical meaning in a physical language understanding system. It is motivated by John Searle's well known (1980) critique of the then-standard and still influential Computational Theory of Mind in cognitive science, which he took to be inadequate because it fails to explain the original or intrinsic meaning characteristic of human mentality. The proposed model combines insights from current philosophy of mind and cognitive neuroscience with a tradition in Western thought whereby the meaning of a word is its signification of a mental concept, and a mental concept is a representation of the mind-external environment causally generated in the mind by the cognitive agent's interaction with that environment. The essence of the proposal is that the meaning of a word is implemented as an activation pattern or, in dynamical systems terms, a stable point attractor in a neural system. The first part of the discussion explains the nature of Searle's objection, the second presents the model, and the third justifies its validity.

Keywords: Lexical meaning, intrinsic meaning, neural dynamical system, artificial neural network modelling, computational theory of language

Introduction

This paper proposes an implementation model for intrinsic lexical meaning in a physical language understanding system. It combines insights from current philosophy of mind and cognitive neuroscience with a tradition in Western thought whereby the meaning of a word is its signification of a mental concept, and a mental concept is a representation of the mind-external environment causally generated in the mind by the cognitive agent's interaction with that environment.The essence of the proposal is that the meaning of a word is implemented as an activation pattern or, in dynamical systems terms, a stable point attractor in a neural system. The first part of the discussion explains the nature of Searle's objection, the second presents the model, and the third justifies its validity.

1. **Motivation**

The black box problem in system identification (Tangirala 2014) builds models of physical systems based on observation of their response to system input: given a box whose internal mechanism is hidden but whose input-output behaviour is observable, what mechanism inside the box generates that behaviour? The answer is that there is an arbitrary number of possible different mechanisms for any given behaviour (Arbib 1987: Ch. 3.2); the only way to know for certain what's in the box is to look inside.

For linguistic meaning the black box is the human head and the input-output behaviour is conversation. The currently-dominant view of what's in the head, the Computational Theory of Mind (CTM; Rescorla 2020), is that it is a Turing Machine whose program is cognition. When the box is opened, however, one looks in vain for the data structures and algorithms of CTM, and finds instead billions of interconnected neurons.

Some have argued that study of the brain by cognitive neuroscience will supplant the theoretical ontology of CTM, but this is not the majority view (Ramsey 2019). Applied to physical systems, the doctrine of emergence in the philosophy of science (O'Connor 2012) addresses the relationship of physics to the "special" sciences, which study objects, properties, or behaviours that emerge from the physical substrate of the natural world. The standard view is that the sciences are related via levels of description (O'Connor 2020) whereby any physical system can be described at an arbitrary number of levels using a theoretical ontology appropriate to each, every level of is explanatorily autonomous with respect to the others subject to the constraint of consistency between and among levels, and selection of any particular level is determined by the research question being asked; the principle of supervenience (McLaughlin and Bennett 2018) says that descriptions of natural systems constitute a hierarchy where the properties at any given level implement those at the level above. For the physical monist (Stoljar 2015), everything is physical and therefore describable using the theoretical ontology of physics, but this does not rule out the ontologies of sciences addressing supervenient phenomena or require their reduction to physics (van Riel and van Gulick 2019; Stoljar 2015) on the grounds that different theoretical ontologies are needed to capture different sorts of regularity in nature.

The present discussion adopts nonreductive physicalism (Stoljar 2015), which in a cognitive science context says that accounts of the structure and operation of mind and brain are separate and autonomous levels of description. It accepts that cognition is ultimately generated by and only by the physical brain, but maintains that this does not preclude the mentalistic ontology of CTM or require its reduction to neuroscience.

A fundamental problem with CTM identified by John Searle in 1980 remains unresolved, however. Searle's initial aim was to counter the claim by "strong" artificial intelligence that it is possible to construct machines with human-level intelligence by basing their design on CTM, but he subsequently extended this to the philosophical foundations of CTM itself, arguing that it cannot in principle offer a complete theory of cognition (Cole 2020). Searle's argument is based on two propositions:

(1) "Intentionality in human beings (and animals) is a product of causal features of the brain".
(2) "Instantiating a computer program is never by itself a sufficient condition of intentionality".

He takes (1) as "an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality", and (2) as something to be established by argument, which he undertakes. The conclusions are that "the explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program", which is "a strict logical consequence" of (1) and (2), and that "any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1. Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain" (all quotations from Searle 1980: 417).

These conclusions assume the validity of (2), which is based on the Chinese Room thought experiment. There is a closed room containing Searle and a list of rules in English for manipulating Chinese orthographic symbols. Chinese speakers outside the room put sequences of these symbols into the room and, using the rules available to him, Searle assembles and outputs sequences of Chinese symbols in response. The people outside interpret the input sequences as Chinese sentences and the output sequences as reasonable responses to them, and on the basis of the room's linguistically coherent input-output behaviour conclude that it understands Chinese. Searle himself, however, knows that the room does not understand Chinese because he, the interpreter and constructor of the sequences, does not understand Chinese, but is only following instructions without knowing what the input and output sequences mean.

The key to understanding the implications for CTM is intentionality (Jacob 2019; Morgan and Piccinini 2017; Neander 2017). The term is related to Latin *intendere*, "to point at, to direct", and was used in medieval European philosophy to refer to the mind's ability to direct its attention to specific mental concepts as well as to things and states of affairs in the mind-external world. In present-day philosophy of mind "intentionality" is used to denote the aboutness of mental states, "the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs" (Jacob 2019). From at least the time of Aristotle (384–322 BC), one way of understanding this aboutness has been via a distinction between conventional signification, where objects are selected as symbols by a community and their referents are determined by agreement, and natural signification, where things by their nature signify independently of conventional agreement, such as smoke signifying fire (Meier-Oeser 2011). Words in natural language are seen as conventional symbols, but the concepts to which they refer are taken to be natural signs of things and events in the world whereby there is a causal nonconventional connection between states of the world and how they are represented in the mind, and consequently a resemblance between the structure of mind-external reality and mind-internal conceptual structure (Moisl 2020). Until the end of the eighteenth century concepts were thought of as mental pictures of the world. In present day cognitive science this view is controversial at best (Pitt 2020; Thomas 2014; Fodor and Pylyshyn 2015), and the dominant view, articulated in the Representational Theory of Mind (RTM; Pitt 2020), has been that intentionality is based on mental representations, that is, physically individuated tokens in the heads of cognitive agents, where each token has a semantic interpretation whose "content" is a state of the world or another head-internal representation and whose form has no necessary resemblance to that which it represents in the world. Representations thereby connect the agent with the mind-external world, and this connection is the foundation on which intentionality is constructed. The notion of mental representation has itself become controversial; for reviews see, for example, (Adams &andAizawa 2017; Bartels 2007; Laurence and Margolis 2019; Margolis and Laurence 2015; Morgan 2014; Morgan and Piccinini 2017; Piccinini 2015; Pitt 2020; Ryder 2009a; Ryder 2009b; Shea 2018). The present discussion accepts the validity of representation in the theoretical ontology of cognitive science.

Searle distinguished two types of intentionality, original and derived, where the locus of original intentionality is the human head and derived intentionality is that which we attribute to physical mechanisms which we believe not to have original intentionality, such as thermostats, whose operation we routinely interpret as wanting to maintain an even temperature but whose structure is too simple for it to have desires. Based on this distinction, his argument is that, with respect to intentionality, a computer is like a thermostat. The Chinese Room is, of course, a computer. Searle is the CPU, the list of English instructions is a program, and the input-output sequences are symbol strings; by concluding that the room understands Chinese, its observers have confirmed the Turing Test (Oppy and Dowe 2016), which says that any device which can by its behaviour convince human observers that it has human-level intentionality must be considered to possess it. Searle knows, however, that the room's intentionality is derived, the implication being that physical computer implementations of CTM models, like thermostats, only have derived intentionality. The intentionality of the symbols manipulated by the algorithm of a CTM model is in the heads and only in the heads of their human designers. When physically instantiated, for example by compilation of a CTM model onto a physical computer, this intentionality is lost: the symbols of the interpreted model cease to be symbolic and become physical tokens which drive the physical causal dynamics of the machine, but intentionality is not a factor in that dynamics. The behaviour of the machine can be interpreted as intentional, just as the behaviour of the Chinese room or of a thermostat can be, but the semantics is derived because the only locus of intentionality is in the heads of observers. A physical computer does not understand what it is doing any more than a vending machine does. It simply pushes physical tokens around, and humans interpret that activity as intentional (Piccinini 2016).

Searle's attack on CTM precipitated a controversy which continues to the present day, but thus far there is no consensus on the validity of his position (Cole 2020). As such, it seems sensible to look for an alternative to the prevailing mainly-philosophical discussion of it. The one adopted here is empirical: assume that Searle is right about the original / derived intentionality distinction and that a physical system can only have original intentionality if intentionality is a causal factor in its operation, and construct models based on these assumptions to see if any useful insights ensue. The present discussion does this by constructing an implementation model for word meaning with original, or as it will henceforth be called, intrinsic intentionality.

2. **The model**

Searle's view that intentionality is generated by and only by the physical brain is the default position in cognitive science. On that view, an obvious approach is to model the neural mechanisms which physically implement cognition (Petersson and Hagoort 2012). This section consequently proposes a model of lexical meaning comprising a collection of interconnected artificial neural networks (ANN) theoretically described as a nonlinear dynamical system which, when physically implemented, has intrinsic intentionality.

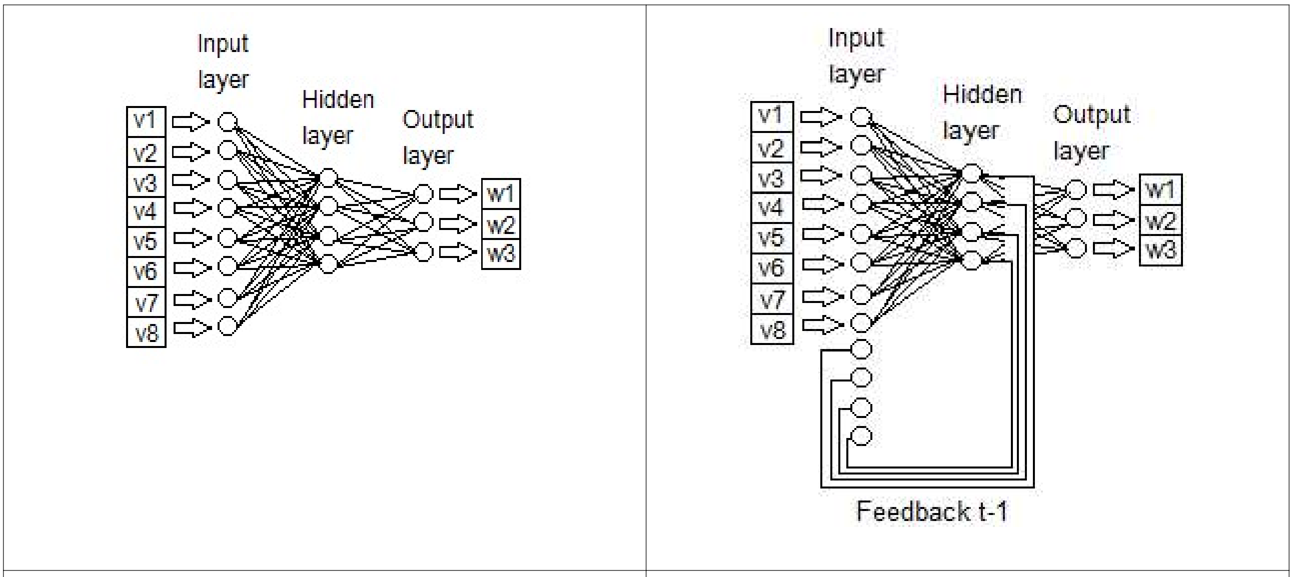
The first part of this section briefly describes the ANNs used in the model, the second motivates their interpretation as dynamical systems, the third specifies the model, and the fourth exemplifies its operation via a simulation.

2.1 **Artificial neural networks**

An ANN (Goodfellow and Bengio 2017) emulates the physical structure and dynamics of biological brains to some approximation; in cognitive science ANNs have been developed under the label "connectionism" (Buckner and Garson 2019). ANN architectures vary, but they all have at least this much in common:

- Structurally they consist of more or less numerous interconnected artificial neurons called "units".
- The units are partitioned into input units that receive signals from an environment, output units that output signals to an environment, and internal "hidden" units that are only accessible via the input units.
- Each unit has connections to other units through which it receives signals. The aggregate of these signals at any time *t* elicits an "activation" that, in the case of the input and hidden units, is propagated along all outgoing connections to other units in the net, and in the case of the output units is made available to the environment.
- A connection between units may be more or less efficient in transmitting signals. The relative efficiency is referred to as "connection strength", and there is typically a significant variation in strength among connections. This variation determines the pattern of unit activations and thereby the network's response to environmental input.
- The variation in connection strength is rarely if ever specified by the designer but is learned by altering the strengths incrementally over time in response to input from the environment.

Two standard architectures are used here: the multilayer perceptron (MLP) and the simple recurrent network (SRN), shown in Figure 1.



a. MLP

b. SRN

Figure 1: ANN architectures

The input and output signals and the input, hidden, and output layers of a physical ANN are mathematically represented as vectors, the connections between layers as matrices, and the propagation of inputs to outputs through physical connections as vector-matrix multiplication. When presented to the input units, the numerical values in an input vector are propagated through the net via the connections and emerge as numerical values in the output units. An MLP is thereby interpretable as a function that maps sets of input vectors to sets of output vectors. The mapping is determined by the values in the input-to-hidden and the hidden-to-output matrices representing the connections between the network layers.

The ANN learns the mapping between sets of input and output vectors using a gradient descent procedure known as backpropagation, for the details of which see (Goodfellow and Bengio 2017). Given a set of vectors V of dimensionality *m* and another W of dimensionality *n*, the MLP training learns a specific mapping from V to W by adjusting the connections between layers such that when v_i ε V is presented at the input layer, w_j ε W appears at the output layer. This is shown in Figure 1 for *m* = 8 and *n* = 3. An SRN augments the MLP with feedback connections among units, so that input to any given unit at time *t* is determined not only by the current input but also by the state of the hidden layer at time *t*-1. An SRN is thereby able to model time-series.

The pattern of MLP hidden layer activations for any given input-output mapping is key to the model here proposed. The reason for this is best seen with reference to a specific type of MLP, the autoassociative MLP or aMLP, shown in Figure 2.

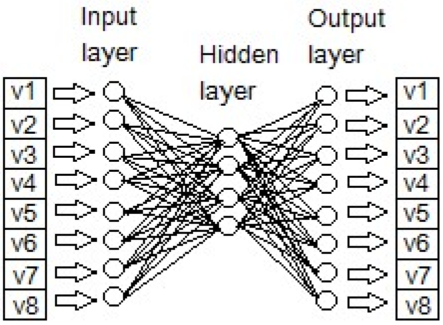


Figure 2: An autoassociative MLP

Input and output are identical, so that after training the aMLP implements the identity function. The hidden layer of a trained aMLP is a representation of its input domain, where "representation" is understood in its etymological sense of a re-presentation of any given form in some different form: in Figure 2 the location of the point specified by the values of the input vector in 8-dimensional space is re-presented as its location in 4-dimensional hidden layer space.

2.2 **Dynamical systems interpretation**

Mathematical dynamical systems theory (Strogatz 2014) is used to model physical systems that change over time. A dynamical model is a triple D = (S,T,*f*), where

- S is an *n*-dimensional state space. Assuming a Euclidean space, the dimensionality is determined by the number of variables used to describe the physical system and each of the *n* orthogonal axes measures a different variable. Let x_1, x_2, ... x_n be variables of some model M. Then the vector of variable values $x_1(t)$, $x_2(t)$ ... $x_n(t)$ is the state of M at time *t*.
- T is a temporal domain, which is an interval of times between t_0 and t_m.
- *f* is a state transition function *f* = S x T -> S which determines the succession of states S_i at times t_i for *i* in the interval 1..*m*, that is, how the state evolves over time.

Geometrically, the state vector is a point in the *n*-dimensional state space, and the change in its values at times t_i describes a trajectory in the space. Depending on how S, T, and *f* are defined, dynamical models can be subcategorized in several ways: S and T can be discrete or continuous, *f* can be linear or nonlinear, and the system as a whole can be autonomous or nonautonomous, where the time evolution of a nonautonomous system is affected by system-external inputs and that of an autonmous system is not. A main focus in the study of dynamical systems is to determine if a system will settle into a stable state over time and, if so, what kind of behaviour that will be. The present discussion is interested in discrete time, continuous space, non-autonomous systems with stable fixed-point attractors.

Why interpret ANNs as dynamical systems? An anlogous question is: why interpret brain physiology and dynamics as a Turing Machine? The answer to both is that it allows the behaviour of a physical object to be understood in terms of existing scientific theory and associated mathematics. In the CTM case this understanding is stated via the theory of computation in terms of data structures and algorithms, and in the dynamical systems case in terms of states and state trajectories. There is an arbitrary number of possible ANN architectures that implement any given function, be it the number and functionality of component subnets, the nature of subnet interaction, and / or the number of units and pattern of connectivity within each component. A dynamical system interpretation generalizes over such variations and allows them to be understood as a class of systems.

2.3 **The model**

As noted, the model presented here is based on a tradition in Western thought whereby the meaning of a word is its signification of a mental concept, and a mental concept is a representation of the mind-external environment causally generated by the cognitive agent's interaction with that environment. An outline history of that tradition from Aristotle to the present day is given in (Moisl 2020); in present-day cognitive science it continues in the attempt to "naturalize" the mind, that is, to see the mind as an aspect of the natural world and therefore theoretically explicable in terms of the natural sciences (Morgan and Piccinini 2017; Papineau 2020). The precursors of naturalism were empiricist philosphers like like Mill (1806-73; Macleod 2016) and scientists like von Helmholtz (1821–1894; Patton 2018) and Mach (1838–1916; Pojman 2019); von Helmholtz stressed the importance of sensory perception of and bodily interaction with the environment in generating a coherent system of mental representation whose structure mirrors that of the environment, and Mach saw human mentality as a teleological dynamical system tending to equilibrium with the environment via sensory and enactive interaction. In the present day, the tradition exists in a variety of disciplines and approaches to the study of mind and language: naturalistic (Rysiew 2020), evolutionary (Bardie and Harms 2020), and teleological (Neander 2012; Neander 2017) epistemology, externalist semantics in philosophy of mind (Lau and Deutsch 2014), evolutionary psychology (Downes 2018) and embodied cognition (Anderson 2003; Barsalou 2010; Wilson and Foglia 2015) in cognitive psychology, and cognitive linguistics (Gärdenfors 2014; Geeraerts and Cuyckens 2012) and conceptual semantics (Jackendoff 2002; Jackendoff 2012) in linguistics.

2.3.1 **Model architecture**

The model, henceforth W, is based on a combination of Searle's requirement that "any mechanism capable of producing intentionality must have causal powers equal to those of the brain*"*, and the tradition just outlined. It is inspired by the Churchlands' neurobiologically grounded State Space Semantics as comprehensively articulated in (Churchland 2012; for critiques see (McCauley 1996; Fodor amd Lepore 1996a; Fodor amd Lepore 1996b), and for defenses (Laakso and Cottrell 2000). They propose a dynamical systems account of how what is known of brain physiology and temporal processing generates cognition. In essence, the Churchland model assumes a structured evironment and, beginning with the neonatal brain, a stimulus from the environment to a sensory modality generates a pattern of neural activation in the brain. With repeated presentation of any given stimulus A, dendritic and synaptic growth maps A to activation in a specific brain location. Cognitively this is learning; from a dynamical systems viewpoint the location is an attractor, and in linear algebraic terms it is a vector that specifies a point in *n*-dimensional space, where *n* is the dimensionality of the vector representation of the physical brain activation. Via the same mechanism, stimuli B, C, D... are mapped to brain locations whose distance from one another is homomorphic with the similarity structure of the stimuli. As learning accumulates for a large number of stimuli, an activation structure emerges in which similar stimuli generate activation in closely adjacent brain locations, and dissimilar stimuli activate relatively more distant locations. Each resulting concentration of activation locations is, in dynamical systems terms, a basin of attraction to which any future stimuli of a similar type are attracted, and in linear algebraic terms is a cluster of points. Over time, for each sensory modality, there emerges a map of basins of attraction / a manifold in *n*-dimensional space. Such maps are modality-specific representations of the external environment. Connections exist not only from sensory input modalities to brain maps, but also between maps and to association areas, the last-mentioned of which are activation areas which learn attractors from the co-activation of the sensory areas to which they are connected. These association areas implement cognitive concepts. The configuration of numerous interconnected sensory and association maps is formed in early life, and their representation of the mind-external environment is the basis for subsequent cognitive activity and learning; as Churchland puts it, the configuration is "the enduring conceptual framework with which [the cognitive agent] will interpret its sensory experience for the rest of its life" (2012: Preface). Interaction with the environment via sensory and motor functions generates state trajectories through the map structure via the connections such that environmental interactions generate trajectories whose relative similarity through the attractor maps is homomorphic with the structure of the perceived and enacted environmental interactions, with gradual modification of map structure and connectivity via dendritic and synaptic growth and atrophy.These trajectories represent the temporal structure of the environment.

An influential development of the Churchlands' original position was that of Laakso and Cottrell (2000), who started with the assumption that "the structure of qualitative content to be reflected in the structure of neural representations at various stages of sensory processing" (Laakso and Cottrell 2000: 50), but proposed that the theoretical representational primitive should be partitioning of the high-dimensional neural state space into activation subspaces, replacing absolute position in the space as originally proposed by the Churchlands: the number of units in the neural model determines the dimensionality of the state space and representations are vectors in the space, but because similar inputs cause similar vectors, activation - clusters form, and these clusters constitute the partition. This idea has subsequently been refined by Shea (2014; 2018), who argues that clusters in state space are the brain's "vehicles of content".

The model proposed in what follows comprises a structure of interacting ANNs in which transduction subnets generate representations of sensory inputs in their hidden layers, and these representations are amalgamated in the hidden layer of an association subnet; the hidden layer of the association subnet for any given word represents its association with inputs from other sensory systems, and that representation is here claimed to implement the intrinsic meaning of the word. Justification of this claim follows in due course.

Figure 3 shows the interconnected MLP, aMLP, and SRN subnets, where numbers of units in each are for illustration only. It has three main components: one to generate representations of acoustic lexical input, one to generate representations of visual input, and one to associate the lexical and visual representations. Only auditory and visual senses are modelled, though others could be added.

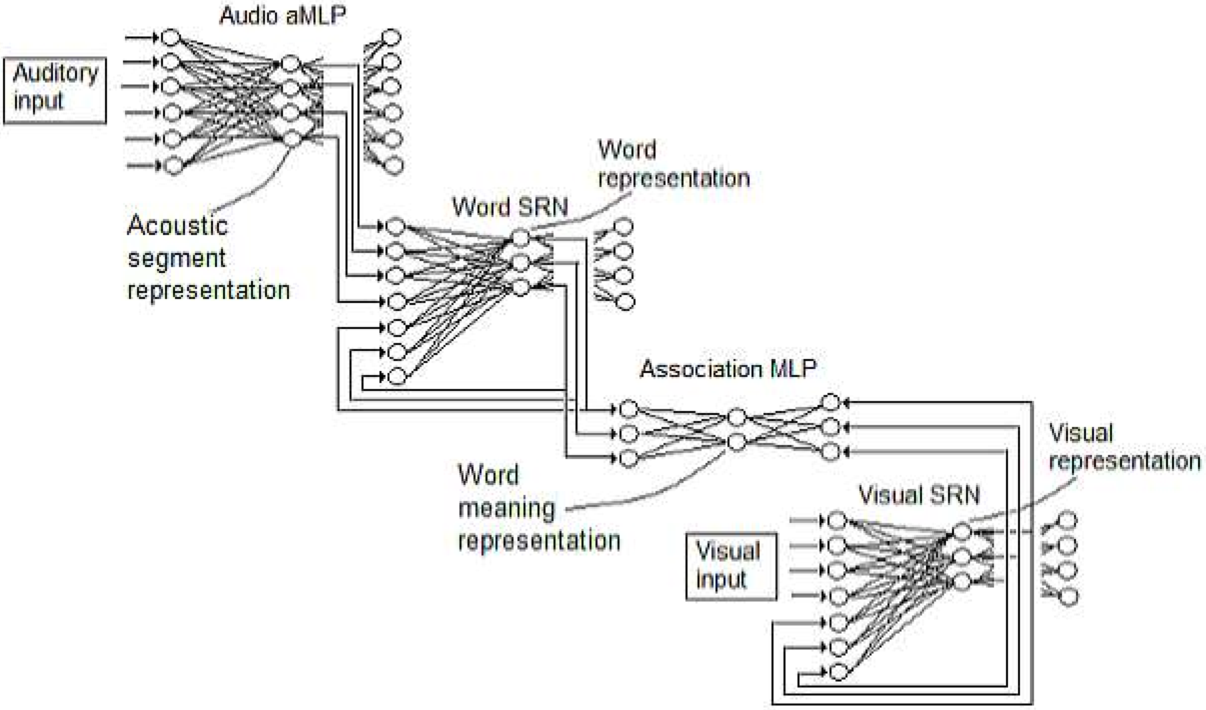


Figure 3: Structure of W

W is trained using a set of of (acoustic, visual) pairs where, for every pair, the acoustic component is a spoken word judged by a human to refer to the visual one.

- Input to the visual subnet is a collection $v_1...v_m$ of visual temporal sequences from an environment. Each sequence v_j is digitally transduced into a sequence of vectorial representations and input to the SRN, where the learning procedure converges on a sequence-final hidden layer activation pattern. This pattern is interpreted as a representation in the above sense: if the sequence-final v_i is a digital activation pattern with *m* components and the hidden layer of the SRN comprises *n* units, where $m \neq n$, then the hidden layer is an alternative re-presentation of $v_{i.}$ In mathematical terms, the hidden layer re-presents the *m*-dimensional manifold as an *n*-dimensional one.
- The auditory subnet is a cascade comprising an aMLP and an SRN. Input is a collection of spoken words $w_1...w_m$ and each w_i is time-sliced into fixed-width phonetic segments. For each word w_i (i) the auditory aMLP sees a temporal sequence of phonetic segments; each phonetic segment is autoassociated, and the hidden layer representation of that segment is sent as input to the word SRN, and (ii) the word SRN processes the sequence of acoustic segment representations for w_i, generating a corresponding sequence of hidden layer activation patterns. The word-final hidden layer pattern is taken to be a representation of w_i.
- The association subnet is an MLP that learns to associate the representation of the acoustic input with the corresponding representation of the visual input. For each (acoustic, visual) pair, the visual representation is the target output for training. Once training is complete, presentation of the acoustic input generates the corresponding visual representation, and the hidden layer is a representation of the mapping from word representation to visual representation.

The correlation in W between states of the world and the words which denote them is an instance of a foundational principle of causal theories of mental content (Adams and Aizawa 2017), which hold that mental content is determined by "normal" conditions, that is, by the way the world typically is. Piccinini and Scarantino (2010; 2011) call this sort of content "natural semantic information", an idea which has its roots in classical antiquity and which, following Dretske (1981) and Fodor (1990), takes the form of mental representations physically caused by probabilistically reliable correlations between real-world events such as smoke and fire.

2.3.2 **Simulation**

A small simulation is presented to exemplify the above process. For tractability, the structure in Figure 3 is simplified.

- Segmentation of continuous acoustic streams is replaced by discrete alphabetic letter sequences, where each letter is a 12 x 12 bitmap, shown in Figure 4, row-wise concatenated to yield a 144-dimensional input vector. The audio aMLP in Figure 3 is replaced by a letter bitmap aMLP.

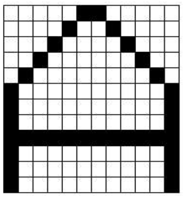


Figure 4: A letter bitmap

- The temporal sequences input to the visual subnet are replaced with static images, and the corresponding SRN with a visual bitmap aMLP. Like the letters, each image is a 12 x 12 binary bitmap row-wise concatenated.

Input is a set of (word, visual) pairs such that there is an intuitively coherent semantic connection between word and image. Assuming *m* (word, visual) pairs, training proceeds by randomly selecting one of the *m* pairs and then presenting them as input to the visual and acoustic subnets respectively. These inputs are propagated through all the subnets, and then the learning algorithm is applied to each. This procedure of selecting a (visual, sentence) pair, processing it, and applying the learning algorithm is iterated until there is no further change in the connections and, consequently, the hidden layer activation patterns in all subnets for any given input have stabilized. Seven words and corresponding images are shown in Figure 5.

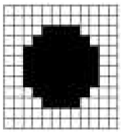
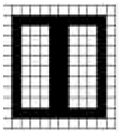
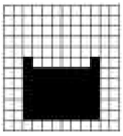
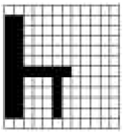
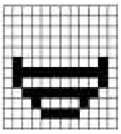
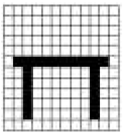
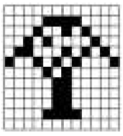
						
1. ball	2. book	3 box	4 chair	5 pot	6 table	7 tree

Figure 5: (Word, image) training pairs

Letter and visual subnet training

The hidden layer generated at each iteration of letter learning was saved in a list and, after training, the vectors in the list were dimensionality reduced using PCA and scatter-plotted. This produced the learning trajectories shown in Figure 6.

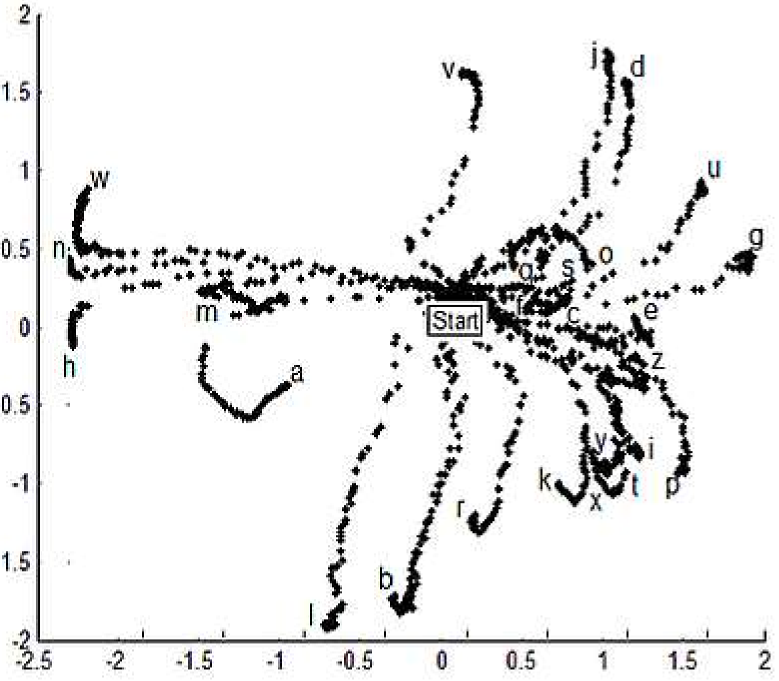


Figure 6: Letter learning trajectories

The training procedure drives each of the inputs from a common initial state through a unique trajectory to a point attractor in the state space, and each attractor is a representation of corresponding bitmap at a unique location in the space. The same was done for the visual subnet, with the result shown in Figure 7.

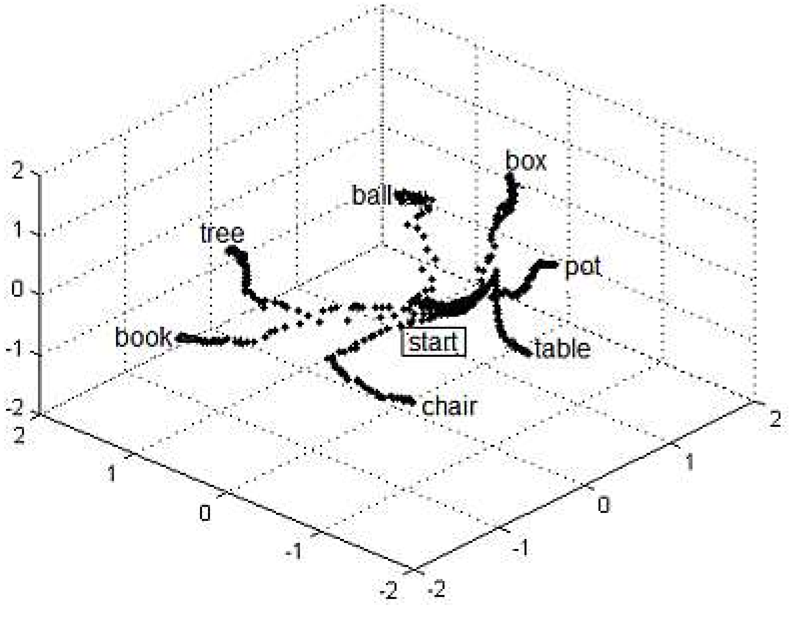


Figure 7: Visual learning trajectories

Word subnet training

Input to the word subnet is a set of sequences of letter representations generated by the letter subnet, where each sequence represents an alphabetic word. For each sequence, the word subnet autoassociates each successive letter representation, and the sequence-final hidden layer activation pattern is taken to represent the word. As above, the hidden layers for each step in this process were saved, dimensionality-reduced for display, and plotted. This is shown in Figire 8, where each word sequence follows a unique trajectory in the space, and each word representation has a unique location.

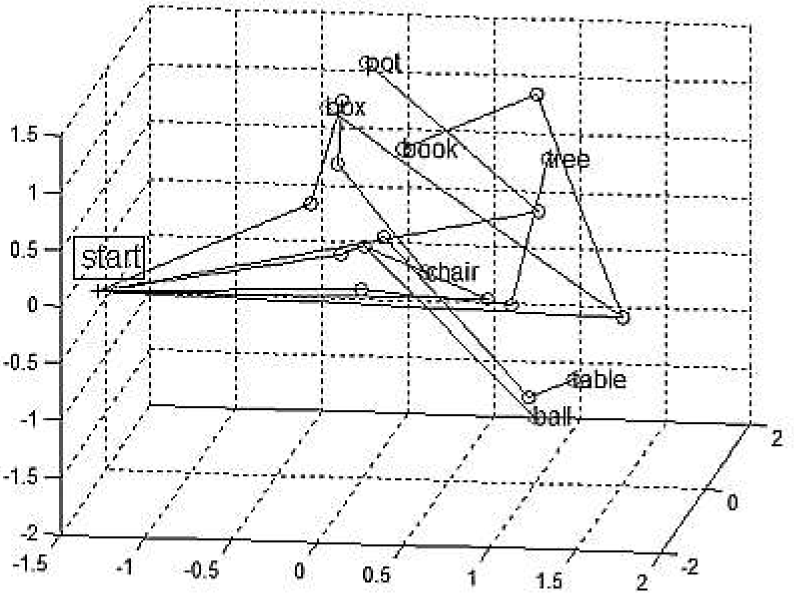


Figure 8: Word learning trajectories

Association subnet training

The association net associates word and visual representations. Training trajectories are shown in Figure 9.

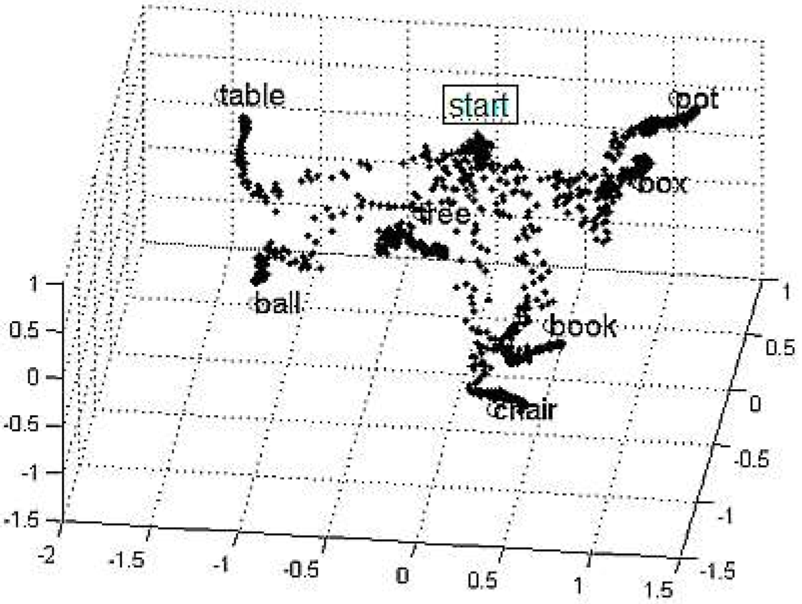


Figure 9: Association trajectories

The representation for each word / visual pair has a unique trajectory ending in a unique location in the state space.

3. **Justification**

If W is to be a valid model of a physical system with intrinsic intentionality, that intentionality must be a causal factor in its dynamics. This section argues that W is such a model. The first part presents a view in current cognitive science of how intentionality is implemented in the brain, and the second shows how W reflects this with reference to word meaning.

3.1 **Neural implementation of intentionality**

The mathematical theory of computation (Sipser 2012) is silent on semantic interpretation of mathematically-specified computational processes. Searle in effect argued that this silence extends to its application in explanation of how intentionality arises in the brain's physical dynamics, an argument strongly supported by recent work on the implementation of mathematically specified computational processes in physical systems (Anderson and Piccinini 2017; Piccinini 2015; Piccinini 2016; Piccinini 2017; Piccinini and Scarantino 2010; Piccinini and Scarantino 2011). How original intentionality arises in brains is not something that the mathematical theory of computation was designed to explain, and interpreting the brain as a physical computational system does not on its own explain its intentionality. Something more is required.

Since the mid-20th century that "something more" has been the representations of RTM. The move away from the traditional "picture" view of mental representation, mentioned earlier, to the RTM one is in semiotic (Chandler 2017) terms a move from icon, which formally resembles what is represents, to symbol, where there is no necessary resemblance between the representation and what is represented. But, as Searle has argued, the conventional semantics of RTM cannot be physically causal in the brain's implementation of intentionality; work outside the standard RTM framework argues for a return to an intuition that underlies the "picture" view but is not identical to it.

Causal theories of mental content explain "how thoughts can be about things...These theories begin with the idea that there are mental representations and that thoughts are meaningful in virtue of a causal connection between a mental representation and some part of the world that is represented. In other words, the point of departure for these theories is that thoughts of dogs are about dogs because dogs cause the mental representations of dogs"; the general principle is that any given symbol "X" means X because "X"s are caused by Xs (Adams and Aizawa 2017; see also Rupert 2008). Proposed causal factors, particularly in teleosemantics (Millikan 2004; Neander 2012; Neander 2017; Thomson and Piccinini 2018) include cognitive development via the individual agent's interaction with a structured physical environment under normal conditions, and the development of cognitive functions by natural selection in the human genome consequent on such experience.

Cognitive neuroscience increasingly provides support for causal theories of mental content via identification of correlations between the various aspects of cognition and brain dynamics in the cognitive agent's interaction with the environment (Boone and Piccinini 2016; Piccinini 2016; Piccinini 2017; Gazzaniga et al. 2019; Piccinini and Bahar 2013; Piccinini and Scarantino 2010; Piccinini and Scarantino 2011; Piccinini and Shagrir 2014; Thomson and Piccinini 2018). Because CTM explains cognition in terms of computation over representations, cognitive neuroscience has been particularly interested in implementation of representations in the brain (Barsalou 2016b; Boone and Piccinini 2016; Thomson and Piccinini 2018; Wilson-Mendenhall et al. 2013). An emerging view is that representations are dynamic neural activation patterns distributed over disparate areas of the brain which are proximately or ultimately based on processes in the brain's sensimotor areas generated by interaction with the environment (Barsalou 2017; Conway and Pisoni 2008; Pulvermũller 2013; Thomson and Piccinini 2018), and that a hierarchy of association areas or "convergence zones" integrates activations from the various sensimotor and other association areas to generate increasingly abstract representations (Anderson 2010; Barsalou 2016a; Barsalou 2016b, Barsalou 2017; Binder 2016; Binder et al. 2005; Binder et al. 2009; Binder et al. 2016; Fernandino et al. 2016; Wilson-Mendenhall et al. 2013). Such association areas provide a plausible implementation mechanism for the question of how sensory input is integrated in the mind so as to generate abstract concepts (Barsalou et al. 2018; Churchland 2012; Feldman 2006; Ryder 2004).

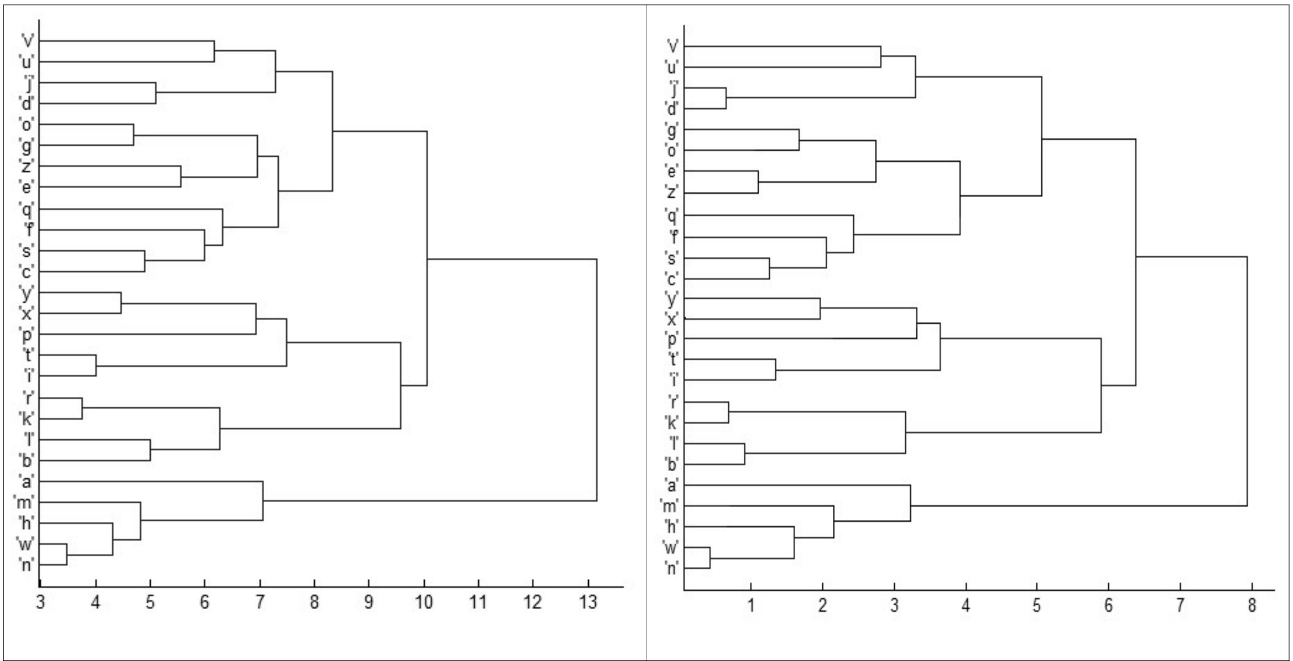
Neurolinguistics, or alternatively biolinguistics (Berwick et al. 2012; Boeckx and Grohmann 2013; Bookheimer 2002; Friederici 2017; Kemmerer 2015; Petersson et al. 2012; Petersson and Hagoort 2012) studies brain processes implementing natural language. Empirical results have shown that the language network is integrated with the general multifunctional organization of brain regions, though it has also been possible to distinguish areas specific to semantic processing and their connection with sensimotor processing (Binder et al. 2009; Friederici 2017; Plebe and de la Cruz 2016). With respect to word meaning (Binder et al. 2009; Fernandino et al. 2016; Garagnani and Pulvermüller 2016; Kemmerer 2015: Chs. 10-12; Pulvermũller 2012;Tomasello et al. 2017), the Grounded Cognition model sees the referents of linguistic expressions as mental representations of mind-external reality, and mental representations as based ultimately on perceptions of that reality mediated by the various motor and perceptual areas; the neural implementation of representations is seen as based on the physical activations of these areas in response to external stimulation and motor interaction with the environment. The closely related Hub-and-Spoke model sees sensimotor areas as physically connected to and integrated in synthetic representations which are the hubs and the sensimotor-specific representations the spokes; the hub integrates cortically distributed sensimotor features of the mental representations to which words refer.

A recurring idea in this and related work is that the structure of the environment is represented as a neurally-implemented model, and that such models can be interpreted in terms of the mathematical concept of homomorphism, "same-formism", a structure-preserving map between two algebraic structures such as vector spaces (Smirnov 2002). Applied to cognition, the idea is that there is a homomorphism between the spatial and temporal structures of the mind-external environment and its representation in the head of the cognitive agent which is causally generated by the agent's interaction with the environment. This idea was proposed in Antiquity (Moisl 2020) and, more recently, by Mach and von Helmholtz, cited above; current examples are (Adams and Aizawa 2017; Bartels 2006; Churchland 2012; Gallistel 1990; Gallistel 2008; Gallistel and King 2009; Gładziejewski and Miłkowski 2017; Garagnani and Pulvermüller 2016; Isaac 2013; Matheson and Barsalou 2018; Morgan and Piccinini 2017; Neander 2017; Piccinini 2018; Piccinini and Bahar 2013; Piccinini and Scarantino 2011; Rescorla 2009; Rupert 2008; Shagrir 2018; Shea 2007; Shea 2014; Shea 2018: Ch.5; Thomson and Piccinini 2018). The relevance of such homomorphic models to present concerns is that they can be understood as implementations of intrinsic intentionality in biological brains because the formal similarity structure of the tokens that causally drive brain dynamics together with the dynamics themselves reflect the similarity structure of mind-external objects and their interactions, and are thereby "about" the mind-external world without involvement of a system-external interpreter.

3.2 **W and homomorphism**

The interpreter-mediated arbitrariness of the connection between representation and what it represents in RTM severs the causal connection between a physical system and the world. W restores it by implementing a homomorphism. Referring to the aMLP in Figure 2, training with respect to some vector *v* configures the connections such that presentation of *v* generates a pattern of hidden layer unit activations specific to *v*, which in turn generates the corresponding output. That pattern is a representation of the 8-dimensional vector in 4-dimensional space. Such a representation is not conventional: its location in 4-dimensional space is caused, via learning, by *v*'s location in 8-dimensional space. In an aMLP, therefore, the hidden layer is a natural signifier in that the form of what is represented causes the form of what represents it; the semantic content of a reduced-dimensionality hidden layer of an aMLP is the higher-dimensional activation pattern which generated it. In a real-world environment the representations learned by an aMLP thereby restore the causal connection between representation and world. Because, moreover, it is an MLP with feedback connections, the case for an SRN is analogous.

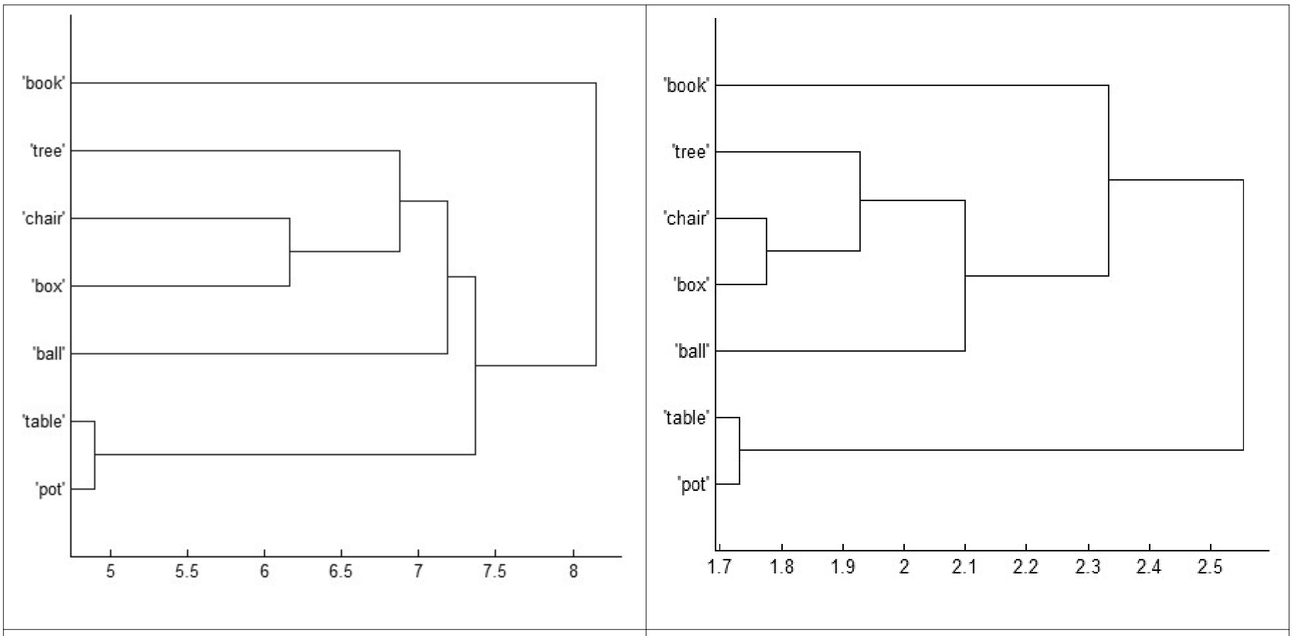
Where an aMLP is trained with two or more input vectors, their similarity relations and those of the hidden layer representations which they generate are homomorphic. This is shown in Figures 10 and 11 by comparative cluster analyses of the letter and visual bitmap vector sets and the corresponding hidden layer vector sets which they generate. The trees are homomorphic; small discrepancies in degree of separation among clusters is attributable to suboptimal selection of network parameters.



a. Letter bitmaps

b. Letter representations

Figure 10: Cluster trees for letter bitmaps and their representations



a: Visual bitmaps

b: Visual representations

Figure 11: Cluster trees for visual bitmaps and their representations

W also implements a homomorphism between letter sequences and their representation, again causally, such that identical sequences generate identical trajectories in the state space and different sequences generate trajectories whose degree of separation in the space reflects the sequence differences. This is not obvious from the trajectories in Figure 8, where word similarity is difficult to judge intuitively, so the set of strings A shown in Figure 13 is used instead.

A brief excursus is required at this point. It is well known that the ability of SRNs to represent sequences is severely limited. A widely used alternative is the Long Short-Term Memory (LSTM), whose sequence-representational capacity is far greater (Hochreiter and Schmidhuber 1997). SRNs were used in the foregoing discussion on account of their intuitive simplicity; LSTMs are rather complex and explanation of how they work would have obscured the overall thrust of the discussion. An SRN would have been inadequte with respect to the strings in Figure 13, so an LSTM is used instead. It is important to understand that this does not undermine the argument being made here: an LSTM is a recurrent ANN that is in principle though not in practice interchangeable with an SRN.

The alphabetic symbols comprising the strings in A were bitmap-encoded and representations derived using an aMLP, as described earlier, so that what the LSTM actually saw was sequences of vectorial representations. At each sequence presentation during training the hidden layer was initialized to a known vector of activation values and each symbol in the sequence was autoassociated. When training was complete the strings were presented to the LSTM in succession, and, for each symbol in each string the hidden layer activation vector was added to a list of hidden layer activations. This list was plotted after reduction to dimensionality-3 using principal component analysis, shown in Figure 12.

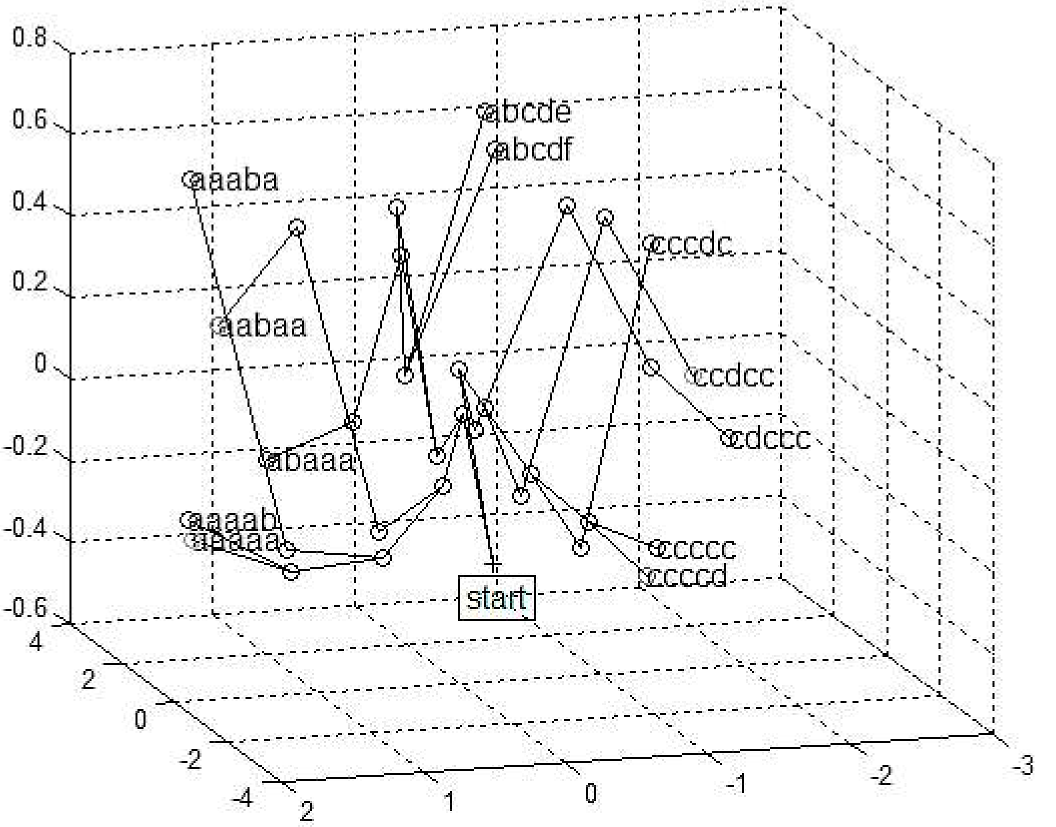


Figure 12: Symbol sequence trajectories

Identical alphabetic sequences follow identical trajectories, as for example the *aa...* ones, but when the sequences differ in a letter they bifurcate and thereafter follow separate trajectories: *aaaaa* bifurcates at *abaaa*, *aabaa*, *aaaba*, and *aaaab*. The same can see seen for for *cc...* and the *abc...* sequences.

Once trained, then, W is a model of a physical system homomorphic with its environment: it (i) generates its own system-internal representations (ii) whose physical forms are determined by that which they represent, (iii) which, for a given environmental domain, represent the similarity structure of the domain and thereby model the domain, and (iv) are causal in the operation of the system. The conclusion is that W's mapping of words to objects in the world is intrinsically intentional because intentionality is causally efficient in the mapping, and that it thereby resolves Searle's problem.

4. **Conclusion**

The discussion has focussed on the specific issue of how a physical system can have intrinsic intentionality with particular reference to word meaning. The argument was that lexical meaning is implemented as a point attractor in a neural dynamical system integrating sensory input. It is not intended as, and obviously cannot serve as, a general model for implementation of word meaning, though the hope is that it can serve as the conceptual basis for one. A full model would, for example, have to incorporate the extensive neuroscientific work on the integration of language and object recognition reviewed in (Plebe and de la Cruz 2016: Ch. 6), as well as addressing such issues as the intentionality of what classical antiquity and medieval scholastic philosophy called universals (Moisl 2020) like "truth" and, more prosaically, the abstract category "human", which have no existence in the mind-external world and cannot therefore generate sensory representations. Nor is it intended as a competitor to the theory of linguistic meaning within the more general computational theory of mind. Cognition may or may not be a Turing-computable function (Copeland 2017; Piccinini 2009; Piccinini 2016; Piccinini 2017), but if it is then a CTM model of it must exist. If that model is to capture the intentionality characteristic of linguistic meaning in the mind, however, there has to be an explanation of how the symbols manipulated by the algorithms of CTM come to have intrinsic intentionality which is causally involved in the generation of cognition, as Searle pointed out. Decades ago, during the classicism / connectionism debate about the appropriate level of cognitive theorizing, Fodor and Pylyshyn (1988) proposed that neural models were best understood as implementation-level accounts of cognitive functions. At the time this struck me as reasonable, and it still does; see further (Piccinini and Craver 2011; Morgan and Piccinini 2017**)**. On this view, interpretation of W in cognitive terms makes it a candidate implementation-level model for intrinsic lexical meaning at the cognitive level.

References

Adams, Fred & Ken Aizawa. 2017. Causal theories of mental content. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, (Summer 2017 Edition). https://plato.stanford.edu/archives/sum2017/entries/content-causal.

Anderson, Michael. 2003. Embodied cognition: a field guide. *Artificial Intelligence* 149. 91–130.

Anderson, Michael. 2010. Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences* 33. 245–266.

Anderson, Neal & Gualtiero Piccinini. 2017. Pancomputationalism and the Computational Description of Physical Systems. https://core.ac.uk/display/78374440.

Arbib, Michael. 1987. *Brains, machines, and mathematics.* 2nd edn. Berlin: Springer.

Barsalou, Lawrence. 2010. Grounded cognition: past, present, and future. *Topics in Cognitive Science* 2. 716-724.

Barsalou, Lawrence. 2016a. On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review* 23. 1122–1142.

Barsalou, Lawrence. 2016b. Situated conceptualization: theory and applications. In *Foundations of embodied cognition. Volume 1: Perceptual and emotional embodiment.* East Sussex: Psychology Press. 11–37.

Barsalou, Lawrence. 2017. What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia* 105. 18-38.

Barsalou, Lawrence, Leo Dutriaux & Christoph Scheepers. 2018. Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B* 373. Issue 1752.

Bartels, Andreas. 2006. Defending the structural concept of representation. *Theoria* 55. 7–19.

Binder Jeffrey. 2016. In defense of abstract conceptual representations. *Psychonomic Bulletin & Review* 23. 1096–1108.

Binder, Jeffrey, Chris Westbury, K. McKiernan, Edward Possing & David Medler. 2005. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience* 17. 905–917.

Binder, Jeffrey, Rutvik Desai, William Graves & Lisa Conant. 2009. Where Is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex* 19. 2767–2796.

Binder, J., Conant, L., Humphries, C., Fernandino, L., Simons, S., Aguilar, M., Desai, R.(2016) Toward a brain-based componential semantic representation, *Cognitive Neuropsychology*, 1–45.

Boeckx, Cedric & Kleanthes Grohmann (eds.). 2013. *The Cambridge handbook of biolinguistics.* Cambridge: Cambridge University Press.

Bookheimer, Susan. 2002. Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience* 25. 151-188.

Boone, Worth & Gualtiero Piccinini. 2016. The cognitive neuroscience revolution. *Synthese* 193. 1509-1534.

Bradie, Michael & William Harms. 2020. Evolutionary Epistemology. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/epistemology-evolutionary.

Buckner, Cameron & James Garson. 2019. Connectionism. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, (Fall 2019 Edition). https://plato.stanford.edu/archives/fall2019/entries/connectionism.

Chandler, Daniel. 2017 *Semiotics. The basics.* 3rd edn. Abingdon: Routledge.

Churchland, Paul. 2012. *Plato's camera. How the physical brain captures a landscape of abstract universals.* Cambridge: MIT Press.

Cole, David. 2020. The Chinese Room Argument. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/chinese-room.

Conway Christopher & David Pisoni. 2008. Neurocognitive basis of implicit learning of sequential structure and its relation to language processing. *Annals of the New York Academy of Sciences* 1145. 113-131.

Copeland, B.Jack. 2017. The Church-Turing thesis, In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/church-turing.

Downes, S. (2018) Evolutionary Psychology. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/evolutionary-psychology.

Dretske, Fred. 1981. *Knowledge and the flow of Information.* Oxford: Blackwell.

Feldman, Jerome. 2006. *From molecule to metaphor. A neural theory of language.* Cambridge: MIT Press.

Fernandino, Leonardo, Jeffrey Binder, Rutvik Desai, Suzanne Pendl, Colin Humphries, William Gross, Lisa Conant & Mark Seidenberg. 2016. Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex* 26. 2018–2034.

Fodor, Jerry. 1990. *A Theory of Content and Other Essays.* Cambridge: MIT Press.

Fodor Jerry & Zenon Pylyshyn. 1988. Connectionism and cognitive architecture. *Cognition* 28. 3–71.

Fodor Jerry & Zenon Pylyshyn. 2015. *Minds without meanings. An essay on the content of concepts.* Cambridge: MIT Press.

Friederici, Angela. 2017. *Language in our brain: The origins of a uniquely human capacity.* Cambridge: MIT Press.

Gallistel, C. Randy. 1990. Representations in animal cognition: An introduction. *Cognition* 37. 1–22.

Gallistel, C. Randy. 2008. Learning and representation. In John Byrne (ed.), *Learning and memory: A comprehensive reference.* Amsterdam: Elsevier. 227–242.

Gallistel, C. Randy & Adam King. 2009. *Memory and the computational brain: Why cognitive science will transform neuroscience.* New York: Wiley.

Garagnani, Max & Friedemann Pulvermüller. 2016. Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *European Journal of Neuroscience* 43. 721-737.

Gärdenfors, Peter. 2014. *Geometry of meaning. Semantics based on conceptual spaces.* Cambridge: MIT Press.

Gazzaniga, Michael, Richard Ivry & G. Mangun. 2019. *Cognitive neuroscience: the biology of the mind.* 5th edn. New York: W. W. Norton.

Geeraerts, Dirk & Hubert Cuyckens. 2012. *Introducing cognitive linguistics.* Oxford: Oxford University Press.

Gładziejewski, Pawel & Marcin Miłkowski. 2017. Structural representations: Causally relevant and different from detectors. *Biology and Philosophy* 32. 337–355

Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2017). *Deep Learning.* Cambridge: MIT Press.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9. 1735-1780.

Isaac, Alistair. 2013. Objective similarity and mental representation. *Australasian Journal of Philosophy* 91. 683–704.

Jackendoff, Ray. 2002. *Foundations of language: Brain, meaning, grammar, evolution.* Oxford: Oxford University Press.

Jackendoff, Ray. 2012. *A user's guide to thought and meaning.* Oxford: Oxford University Press.

Jacob, Pierre. 2019. Intentionality, In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/intentionality.

Kemmerer, David. 2015. *Cognitive Neuroscience of Language.* Hove, UK: Psychology Press.

Laakso, Aarre & Garrison Cottrell. 2000. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology* 13. 47–76.

Lau, Joe & Max Deutsch. 2014. Externalism about mental content. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition). https://plato.stanford.edu/archives/fall2019/entries/content-externalism.

Macleod, Christopher. 2016. John Stuart Mill, In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/mill.

Margolis, Eric & Stephen Laurence. 2015. *The conceptual mind: New directions in the study of concepts*, Cambridge: MIT Press.

Margolis, Eric & Stephen Laurence. 2019. Concepts, In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition). https://plato.stanford.edu/archives/sum2019/entries/concepts.

Matheson, Heath & Lawrence Barsalou. 2018. Embodiment and grounding in cognitive neuroscience. In John Wixted (ed.), *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience.* 4th edn. Hoboken NJ: Wiley.

McLaughlin, Brian & Karen Bennett. 2018. Supervenience, In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition). https://plato.stanford.edu/archives/win2018/entries/supervenience.

Meier-Oeser, Stephan. 2011. Medieval semiotics. In Edward Zalta (ed.),*The Stanford Encyclopedia of Philosophy*, (Summer 2011 Edition). https://plato.stanford.edu/archives/sum2011/entries/semiotics-medieval.

Millikan, Ruth. 2004 *Varieties of meaning.* Cambridge: MIT Press.

Moisl, Hermann. 2020. Intrinsic intentionality and linguistic meaning: an historical outline. In Emmerich Kelih & Reinhard Köhler (eds.), *Words and numbers. In memory of Peter Grzybek (1957-2019).* Lüdenscheid: RAM Verlag.

Morgan, Alex. 2014. Representations gone mental. *Synthese* 191. 213-244.

Morgan, Alex & Gualtiero Piccinini. 2017. Towards a cognitive neuroscience of intentionality. *Minds and Machines* 28. 119-139.

Neander, Karen. 2012. Teleological theories of mental content. In Edward Zalta (ed.),*The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition). https://plato.stanford.edu/archives/spr2018/entries/content-teleological.

Neander, Karen. 2017. *A mark of the mental: In defense of informational teleosemantics.* Cambridge: MIT Press.

O'Connor, Timothy. 2020 Emergent Properties. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/properties-emergent.

Oppy, Graham & David Dowe. 2016. The Turing test. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). URL = <https://plato.stanford.edu/archives/spr2019/entries/turing-test/>.

Papineau, David. 2020. Naturalism, In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/naturalism.

Patton, Lydia. 2018. Hermann von Helmholtz. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition). https://plato.stanford.edu/archives/win2018/entries/hermann-helmholtz.

Petersson, K. , Folia, V., Hagoort, P. (2012) What artificial grammar learning reveals about the neurobiology of syntax, *Brain and Language* 120, 83-95.

Petersson, Karl & Peter Hagoort. 2012. The neurobiology of syntax: beyond string sets. *Philosophical Transactions of the Royal Society B* 367. 1971-1983.

Piccinini, Gualtierio. 2009. Computationalism in the philosophy of mind. *Philosophy Compass* 4. 515–532.

Piccinini, Gualtierio. 2015. *Physical computation. A mechanistic account.* Oxford: Oxford University Press.

Piccinini, Gualtierio. 2016. The computational theory of cognition. In Vincent Müller (ed.), *Fundamental issues in artificial intelligence*. New York: Springer.

Piccinini, Gualtierio. 2017. Computation in physical systems. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, (Summer 2017 Edition). https://plato.stanford.edu/archives/sum2017/entries/computation-physical systems.

Piccinini, Gualtierio. 2018. Computation and representation in cognitive neuroscience. *Minds and Machines* 28. 1-6.

Piccinini, Gualtierio & Sonya Bahar. 2013. Neural computation and the computational theory of cognition. *Cognitive Science* 37. 453-488.

Piccinini, Gualtierio & Carl Craver. 2011. Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese* 183. 283–311.

Piccinini, Gualtierio & Andrea Scarantino. 2010. Computation vs. information processing: why their difference matters to cognitive science. *Studies in the History and Philosophy of Science* 41. 237-246.

Piccinini, Gualtierio & Andrea Scarantino. 2011. Information processing, computation, and cognition. *Journal of Biological Physics* 37. 1-38.

Piccinini, Gualtierio & Oron Shagrir. 2014. Foundations of computational neuroscience. *Current Opinion in Neurobiology* 25. 25-30.

Pitt, David. 2020. Mental Representation. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/mental-representation.

Plebe, Alessio & Vivian de la Cruz. 2016. *Neurosemantics. Neural processes and the construction of linguistic meaning.* Berlin: Springer.

Pojman, Paul. 2019. Ernst Mach. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/ernst-mach.

Pulvermũller, Friedemann. 2012. Meaning and the brain: The neurosemantics of referential, interactive, and combinatorial knowledge. *Journal of Neurolinguistics* 25. 423-459.

Pulvermũller, Friedemann. 2013 How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences* 17. 458-70.

Ramsey, William. 2019. Eliminative Materialism. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/materialism-eliminative.

Rescorla, Michael. 2009. Cognitive maps and the language of thought. *British Journal for the Philosophy of Science* 60. 377-407.

Rescorla, Michael. 2020. The computational theory of mind. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition). https://plato.stanford.edu/archives/spr2020/entries/computational-mind.

Rupert, Robert. 2008. Causal Theories of Mental Content. *Philosophy Compass* 3. 353–380.

Ryder Dan. 2004. SINBAD neurosemantics: A theory of mental representation. *Mind and Language* 19. 211–240.

Ryder, Dan. 2009a. Problems of representation I: Nature and role. In John Symons & Paco Calvo (eds.), *The Routledge Companion to Philosophy of Psychology.* London: Routledge.

Ryder, D. (2009b) Problems of representation II: naturalizing content, In: J. Symons & P. Calvo (Eds.) *The Routledge Companion tp Philosophy of Psychology*, London: Routledge, 251-279.

Rysiew, Patrick. 2020. Naturalism in epistemology. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/epistemology-naturalized.

Searle, John. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3. 417–457.

Shagrir, Oron. 2018. The brain as an input–output model of the world. *Minds and Machines* 28. 53-75.

Shea, Nicholas. 2007. Consumers need information: Supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research* 75. 404–435.

Shea, Nicholas. 2014. Exploitable isomorphism and structural representation. *Proceedings of the Aristotelian Society* 114. 123-144.

Shea, Nicholas. 2018. *Representation in cognitive science.* Oxford: Oxford University Press.

Sipser, Michael. 2012. *Introduction to the theory of computation.* 3rd edn. Course Technology.

Smirnov, D. 2002. Homomorphism. In *Encyclopedia of Mathematics*. http://encyclopediaofmath.org/index.php?title=Homomorphism&oldid=47265.

Stoljar, Daniel. 2015. Physicalism. In Edward Zalta (ed.),*The Stanford Encyclopedia of Philosophy*, (Winter 2017 Edition). https://plato.stanford.edu/archives/win2017/entries/physicalism.

Strogatz, Steven. 2014. *Nonlinear dynamics and chaos.* 2nd edn. Boulder CO: Westview Press.

Tangirala, Arun. 2014. *Principles of system identification: Theory and practice.* Boca Raton: CRC Press.

Thomas, Nigel. 2014. Mental Imagery. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition). https://plato.stanford.edu/archives/sum2019/entries/mental-imagery.

Thomson, Eric & Gualtiero Piccinini. 2018. Neural representations observed. *Minds and Machines* 28. 191-235.

Tomasello, Rosario, Max Garagnani, Thomas Wennekers & Friedemann Pulvermüller. 2017. Brain connections of words, perceptions and actions: A neurobiological model of spatio-temporal semantic activation in the human cortex. *Neuropsychologia* 98. 111-129.

van Riel, Raphaeel & Robert van Gulick. 2019. Scientific reduction. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/scientific-reduction.

Wilson, Robert & Lucia Foglia. 2015. Embodied cognition. In Edward Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition). https://plato.stanford.edu/archives/spr2017/entries/embodied-cognition.

Wilson-Mendenhall, Christine, W. Kyle Simmons, Alex Martin & Lawrence Barsalou. 2013. Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience* 25. 920–935.