Dynamical systems implementation of intrinsic sentence meaning

Hermann Moisl (hermann.moisl49@gmail.com)
Newcastle University https://orcid.org/0000-0002-5911-0373

Research Article

Keywords: Intentionality, sentence meaning, neural dynamical system, artificial neural network modelling, computational theory of language

Posted Date: December 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1214916/v1

Version of Record: A version of this preprint was published at Minds and Machines on January 23rd, 2022. See the published version at https://doi.org/10.1007/s11023-022-09590-1.

Abstract

This paper proposes a model for implementation of intrinsic natural language sentence meaning in a physical language understanding system, where 'intrinsic' is understood as 'independent of meaning ascription by system-external observers'. The proposal is that intrinsic meaning can be implemented as a point attractor in the state space of a nonlinear dynamical system with feedback which is generated by temporally sequenced inputs. It is motivated by John Searle's well known (1980) critique of the then-standard and currently still influential Computational Theory of Mind (CTM), the essence of which was that CTM representations lack intrinsic meaning because that meaning is dependent on ascription by an observer. The proposed dynamical model comprises a collection of interacting artificial neural networks, and constitutes a radical simplification of the principle of compositional phrase structure which is at the heart of the current standard view of sentence semantics because it is computationally interpretable as a finite state machine.

Introduction

This paper proposes a model for implementation of intrinsic natural language sentence meaning in a physical language understanding system, where 'intrinsic' is understood as 'independent of meaning ascription by system-external observers'. The proposal is that intrinsic meaning can be implemented as a point attractor in the state space of a nonlinear dynamical system with feedback which is generated by temporally sequenced input. It is motivated by John Searle's well known (1980) critique of the then-standard and currently still influential Computational Theory of Mind (CTM), the essence of which was that CTM representations lack intrinsic meaning because that meaning is dependent on ascription by an observer.. The proposed dynamical model comprises a collection of interacting artificial neural networks, and constitutes a radical simplification of the principle of compositional phrase structure which is at the heart of the standard view of sentence semantics because it is computationally interpretable as a finite state machine.

The discussion is in three main parts. The first part motivates the model with reference to Searle's distinction of derived and original intentionality, the second describes the model and argues that it implements intrinsic sentence meaning, and the third justifies its finite state architecture.

1. Motivation

Humans intuitively feel that they possess a head-internal meaningfulness, that is, an awareness of the self and its relationship to the perceived world which is independent of interpretation of one's behaviour by observers. Searle (1980) argued that the dominant theoretical framework in cognitive science, the Computational Theory of Mind (CTM; Rescorla, 2020), is inadequate because, as currently formulated, it is incapable of explaining how the human mind comes to possess this head-internal meaningfulness - what Searle called original intentionality, a position which he defended and elaborated in a series of subsequent publications (references in Cole, 2020).

The philosophical concept of intentionality (Jacob 2019; Morgan & Piccinini, 2018; Neander, 2017) is central to Searle's critique of CTM. The term is etymologically 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 and to things and states of affairs in the mind-external world (Moisl, 2020). 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).

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 have good reason to believe do not have original intentionality, such as thermostats, whose operation is routinely interpreted by humans as wanting to maintain an even temperature but whose structure is too simple for it to have desires. His argument is that, with respect to intentionality, a computer is like a thermostat. The argument is based on his well known 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 sentences in Chinese whose meaning they understand and the output sequences as reasonable responses to them, and on the basis of the room's conceptually 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 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 & Dowe, 2016), which says that any device which can by its behaviour convince 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 can be, but the semantics is derived because the only locus of intentionality is in the heads of observers. Put simply, a physical computer does not understand what it is doing any more than a vending machine does. It only pushes physical tokens around, and humans interpret that activity as intentional (Piccinini, 2016).

Searle's position remains controversial after four decades (Cole, 2020). Given this lack of consensus, it seems useful to introduce some empirical evidence into the thus far predominantly philosophical debate. The present discussion does this by assuming the validity of Searle's position and constructing a model based on that assumption to see if any useful insights ensue.

2. The Model

2.1 Methodology

i. *Levels of description*

The black box problem in system identification (Tangirala, 2014) builds models of physical systems based on observation of responses 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 input-output behaviour (Arbib, 1987, ch. 3.2); the only way to know for certain what's in the box is to look inside.

Applied to black boxes in general, the doctrine of emergence in the philosophy of science (O'Connor, 2020) 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 whereby any physical system can be described at an arbitrary number of levels using a theoretical ontology appropriate to each, every level 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 & 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 in the natural world 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 & van Gulick, 2019; Stoljar, 2015) on the grounds that different theoretical ontologies are needed to capture different sorts of regularity in nature.

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). The alternative adopted here is 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 human cognition is implemented 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. The present discussion focuses on the implementation level: how intrinsic sentence meaning can be implemented in a physical system.

ii. *Meaning*

Proposal of an implementation model implies clarity about what is being implemented. 'Meaning' is understood, and its theoretical characterization is approached, in a variety of ways, for an overview of which see (Speaks, 2021). Speaks distinguishes 'logical' approaches in the tradition of Frege, where meaning is seen as semantic interpretation of symbols in an abstract formal system, and 'foundational' approaches which focus on the mechanism of semantic interpretation; foundational approaches are subcategorized into 'use' theories such as those of Grice, and 'mentalistic' ones which relate linguistic meaning to the structure of cognition. The present discussion takes the mentalistic approach, and more specifically adopts the tradition in Western thought (Moisl, 2020) ranging from Aristotle to theories of mental content in present-day linguistics and cognitive science more generally (Adams & Aizawa, 2017) that 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. In recent times this tradition has continued in attempts 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 & Piccinini, 2018; 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, evolutionary, and teleological epistemology (Rysiew, 2020; Bradie & Harms, 2020; Neander, 2017 respectively), externalist semantics in philosophy of mind (Lau & Deutsch, 2014), evolutionary psychology (Downes, 2018) and embodied cognition in cognitive psychology (Anderson, 2003; Barsalou, 2010; Wilson & Foglia, 2015), cognitive linguistics (Gärdenfors, 2014; Geeraerts & Cuyckens, 2012) and conceptual semantics (Jackendoff, 2002, 2012) in generative linguistics.

2.2 Context

Given the foregoing comments, the problem is how to incorporate original, or as it will henceforth be called, intrinsic intentionality into cognitive and more specifically linguistic models such that, when they are physically implemented, the intentionality is causally efficient in their physical behaviour.

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*" (Adams & Aizawa, 2017; see also Rupert, 2008). Proposed causal factors, particularly in teleosemantics (Millikan, 2004; Neander, 2017; Thomson & Piccinini, 2018) include cognitive development via the individual agent's interaction with a structured physical environment under normal conditions.

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 & Piccinini, 2016; Piccinini, 2016, 2017; Gazzaniga et al., 2019; Piccinini & Bahar, 2013; Piccinini & Scarantino, 2010, 2011; Piccinini & Shagrir, 2014; Thomson & 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 & Piccinini, 2016; Thomson & 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 & Pisoni, 2008; Pulvermüller, 2013; Thomson & 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, 2016b, 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; Ryder, 2004).

Neurolinguistics, or alternatively biolinguistics (Boeckx & Grohmann, 2013; Friederici, 2017; Kemmerer, 2015; Petersson et al., 2012; Petersson & 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 & de la Cruz, 2016). With respect to word meaning (Binder et al., 2009; Fernandino et al., 2016; Garagnani & Pulvermüller, 2016; Kemmerer, 2015 Chs. 10-12; Pulvermüller, 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 mathematical homomorphism, that is, of 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 their 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 & Aizawa, 2017; Bartels, 2006; Churchland, 2012; Gallistel, 1990, 2008; Gallistel & King, 2009; Gładziejewski & Miłkowski, 2017; *Garagnani & Pulvermüller, 2016;* Isaac, 2013; Matheson & Barsalou, 2018; Morgan & Piccinini, 2017; Neander, 2017; Piccinini, 2018; Piccinini & Bahar, 2013; Piccinini & Scarantino, 2011; Rescorla, 2009; Rupert, 2008; Shagrir, 2018; Shea, 2007, 2014, 2018 Ch.5; Thomson & Piccinini, 2018).

The relevance of homomorphic implementation-level models to the present discussion 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.

2.3 Model architecture

The model described in this section maps representations homomorphic with linguistic input from the environment to representations homomorphic with visual input from the environment, and argues that a function associating these representations implements intrinsic intentionality. On the assumption that the brain and only the brain implements human cognition, and given the neuroscientific evidence for homomorphism with the environment outlined above, it adopts Searle's position that '*any mechanism capable of producing intentionality must have causal powers equal to those of the brain'.* It is inspired *mainly* by Ryder's neurosemantics (Ryder, 2004) and Churchland's neurobiologically grounded account of cognition as articulated in (Churchland 2012), and comprises the collection of interacting artificial neural networks (ANN) shown in Figure 1; numbers of units are for illustration only, and the component subnets are standard Multilayer Perceptrons (MLP) and Simple Recurrent Networks (SRN), for the architecture and training of which see (Haykin, 2008) and (Goodfellow & Bengio; 2017). For convenience, the model is henceforth referred to as 'S'.

Training input is a set of (sentence, visual) pairs such that there is an intuitively coherent semantic connection between what the visual configuration shows and what the phonetic sequence says - for example, a visual image of a cat running down the street would be paired with the spoken counterpart of

'The cat is running down the street'. This kind of correlation between the state of the world and what is said about it is both intuitively plausible and also a foundational principle of the causal theories of mental content, which hold that mental contents are determined by 'normal' conditions, that is, by the way the world typically is (Adams & Aizawa, 2017).

Assuming *m* (sentence, visual) pairs, each of length *n*, training proceeds by randomly selecting one of the *m* pairs and then presenting them as input to the visual and acoustic subnets respectively in synchronized temporal succession from *t0* to *tn*. The inputs are propagated through all the subnets, and then the training algorithm is applied to each. This procedure of selecting a (sentence, visual) pair, processing it, and applying the training 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.

Training of the sentence component

Every sentence is a discrete sequence of time-sliced acoustic segments, and each segment is represented as a vector. Such a vector sequence is assumed to be partitioned by word boundaries so that, for any given sentence, what the model sees in the course of training is a sequence of words each of which consists of a sequence of phonetic segment vectors.

- The auditory subnet MLP autoassociates each phonetic segment vector. Each autoassociation generates a configuration of activations in the MLP's hidden layer, and this is sent as input to the word subnet SRN.
- The word subnet SRN autoassociates each successive hidden layer configuration from the auditory subnet, and, as with the auditory subnet, each autoassociation generates an activation configuration in the SRN's hidden layer. The word-final hidden layer configuration is sent as input to the sentence SRN.
- The sentence subnet SRN autoassociates each successive hidden layer configuration from the word subnet, and, as before, each autoassociation generates an activation configuration in the SRN's hidden layer. The sentence-final hidden layer configuration is sent as input to the association MLP.

Training of the visual subnet

Each visual is a static bitmap representation of some scene in the natural world as it might be perceived by a human. The bitmaps are row-wise concatenated into vectors, and the vectors are autoassociated by the visual MLP, generating a hidden layer activation configuration for each bitmap vector.

Training of the association subnet

For every (acoustic, visual) training pair, the sentence-final hidden later configuration of the sentence SRN is input to the association subnet MLP, the hidden layer configuration of the visual subnet is the target output, and training associates the two.

When trained, input of a sentence from the training set generates the corresponding visual scene in the output units of the visual MLP.

A small simulation is presented to exemplify the training and operation of S. For tractability, the structure shown in Figure 1 is simplified so that segmentation of continuous acoustic streams is replaced by discrete alphabetic letter sequences, where each letter is a 12 x 12 bitmap, shown in Figure 2, row-wise concatenated to yield a 144-dimensional input vector. The audio MLP in Figure 1 is replaced by a letter bitmap MLP.

The visual scenes and corresponding sentences used in the simulation are shown in Figure 3.

Scene and letter subnet training

Scene and letter vectors were randomly presented and, for each iteration, the hidden layer activations during training were saved so that their evolution over the training iterations could be observed. These were dimensionality-reduced for graphical display using PCA and plotted in Figures 4 and 5; the numbers in Figure 4 refer to the scenes in Figure 3. Only the trajectories for the first four scenes are shown because thereafter the graphic becomes too dense to be interpretable.

In dynamical systems terms, the training procedure over time drives each of the inputs from a random initial state through a unique state space trajectory to a point attractor in the space, and each attractor is a vector at a unique location. The letter training trajectories in Figure 5 are analogous; only the first four letters are shown because, as above, the graphics become too dense for legibility thereafter. As with the visual scenes, each trajectory ends at a unique location in the state space.

Word and sentence subnet training

Each of the individual autoassociations in the sequences processed by the word and sentence SRNs generates a hidden layer training trajectory, but these are too numerous to be exhaustively shown. The trajectories behave like those just described: each goes from a random initial state through a unique trajectory to a unique point attractor.

Association subnet training

The association MLP associates the representations of the input sentences generated in the hidden layer of the sentence net with the representations of the visual scenes generated in the hidden layer of the visual net, using the sentence representation as input and the visual representation as target output. Training trajectories are shown in two dimensions for clarity in Figure 6.

Once trained, S maps the sentences to the corresponding visual scenes in the training set. Figure 7 shows the output of the visual subnet in response to the letter sequences comprising the sentences presented at the letter subnet.

2.4 Interpretation

The key to substantiating the claim that the model implements intrinsic intentionality is the representational capacity of the hidden layers in the subnets. This is clearest with reference to the variant of the MLP, the autoassociative MLP (aMLP), used for the acoustic and visual input subnets of Figure 1 and shown in Figure 8.

The input and target output of an aMLP are identical, so that after training the aMLP implements the identity function, that is, presentation of any vector v on which the aMLP is trained results in v as output. 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 8 the location of the point specified by the values of the input vector in 8-dimensional vector space is re-presented as its location vector in 4-dimensional space.

For an MLP with a specific architecture in terms of the numbers of input-output and hidden units and connection initialization values, the hidden layer representations for an identity mapping of a set of vectors V = {v1, v2...vn) have the following characteristics of particular relevance to present concerns.

- The representation for each vi ε V is causally generated by the training procedure and, in a trained network, is causal in mapping input to output. The connection values are adjusted so that presentation of vi generates the activation pattern in the hidden layer, which in turn generates a duplicate of vi.
- The representation for each vi ε V is unique among the representations for all the vectors in V.
- The connection between each vi and it representation is nonarbitrary: any alteration to the hidden layer by external intervention will compromise the mapping to some degree.

Once trained, the hidden layer activation patterns in the acoustic/letter and visual aMLPs are representations in this sense.

Where an aMLP is trained on a set V containing two or more vectors, the similarity relations among hidden layer vectors in a trained network mirror those of the component vectors of V. This is shown in Figures 9 and 10 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, and, in intepreting the tree diagrams, it has to be kept in mind that subtrees can be rotated about their horizontal axes without affecting the structure.

There is also a homomorphism between the similarity structures of the letter and word sequences on the one hand and the hidden layer trajectories which they respectively generate on the other. Sample trajectories for word-letter sequences from the training set are shown in Figure 11. These were generated by inputting letter sequences for each word in the training set into the letter subnet of the trained model, saving the hidden layer vectors thereby generated in the word SRN, and then 3-D plotting the dimensionality-reduced list of vectors.

Plotting all 25 word trajectories renders the graphic unreadable, so only a few samples are shown. Each letter sequence traces a unique path through the space, and each ends at a unique location.

Figure 12 shows the trajectories for all 6 sentences of the training set, and here too each sentence traces a unique path through the space and ends at a unique location.

It is not clear from Figure 11 that the relative similarity structure of the letter trajectories is homomorphic with the relative similarities of the letter sequences which constitute the words, or from Figure 12 that the relative similarity structure of the word trajectories is homomorphic with the similarities of the word sequences which constitute the sentences, because the relative similarities of words and sentences are not obvious. A string set whose relative similarties are obvious, shown in Figure 13, is therefore used to justify the claim to homomorphism.

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 & Schmidhuber, 1997). SRNs were used in the foregoing discussion on account of their intuitive simplicity, whereas LSTMs are rather complex and explanation of how they work would have obscured the overall thrust of the discussion. For the strings in Figure 13 an SRN would have been inadequate, so an LSTM is used instead. It is important to understand that this does not compromise the argument being made here: an LSTM is a recurrent artificial neural network that is in principle though not in practice interchangeable with an SRN.

An autoassociative LSTM was trained on the strings in Figure 13, and, when training was complete, these were presented to the LSTM in succession, saving the hidden layer vector for each in a list. After PCA dimensionality reduction the list was plotted, as shown. 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.

Finally, there is a homomorphism in the mapping from sentence to visual scene in the association subnet. After training, the sentences and corresponding visuals were presented to the complete net, and the hidden layer configurations for each input pair was saved, dimensionality-reduced, and plotted as shown in Figure 14. The association subnet's hidden layer homomorphically represents the set of input sentences.

In summary, then:

(i) S causally generates its own system-internal representations of external environmental input.

(ii) The physical form of these representations is determined by that which they represent.

(iii) For a given environmental domain, the representations are homomorphic with the similarity structure of the domain and thereby model it.

(iv) The representations are causal in the input-output behaviour of the system.

It was stated earlier that the relevance of homomorphic implementation-level 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. Assuming the validity of this, the conclusion is that S's mapping of sentences to visual states of the world is a model of a physical system that implements intrinsic intentionality, and that it resolves Searle's problem.

Does S thereby implement sentence meaning? The meaning of a linguistic expression in cognitive models of meaning (Speaks, 2021) 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, as noted. The association subnet of S is a function whose domain is sentence representations and whose range is visual representations generated by the environment. On that criterion, S implements sentence meaning.

3. Finite State Architecture

S has the purely sequential dynamics of a finite state automaton (FSA), which appears to disqualify it as a viable natural language model. This section first looks at finite state architecture and how S is an instance of it, then at the nature of the problems it poses, and finally at the resolution of these problems.

3.1 Finite state architecture

Given arbitrary sets A and B, a function *f* is a subset of the Cartesian product C of A x B such that, for every pair (*a,b*) ∈ C, *a* ∈ A is uniquely associated with *b* ∈ B. An algorithm is said to compute *f* if, given the first component *a* of a pair (*a,b*) ∈ C, it returns the second component *b*, that is, if it generates the set of pairs that constitutes *f*. If the physical inputs and outputs of an artificial neural network N are interpreted as A and B, and if N's physical input-output behaviour is consistent with C under that interpretation, then N can be interpreted as a computational system.

The theory of computation studies the classes of function that can be computed by various types of automata (Hopcroft et al, 2000; Sipser, 2012). Under unboundedness assumptions analogous to those made in automata theory, the ANN architectures used in S have been shown to be Turing-equivalent (Siegelman & Sontag, 1995). ANNs that are bounded in terms of the number of units or the precision of activation functions and connection strengths are computationally equivalent to finite state automata and therefore limited in the range of functions they can compute; because all physically-implemented ANNs are necessarily bounded, they are all limited to finite state computational power. This does not

compromise ANNs relative to standard automata, however, because the same applies to the latter - every implemented Turing machine is equivalent to a finite state machine. In real-world applications automata are given sufficient memory to compute the finite subset of the required function in practice, whereas ANNs are given more units and/or higher-resolution activation functions and connection strengths to achieve the same end.

For every computable function *f*, therefore, there exists an ANN architecture that computes a bounded subset of *f*. Where *f* is a linguistic function, an ANN architecture with feedback connections such as an SRN is a dynamical system which a sequence of linguistic inputs drives through a state space trajectory, thereby implementing an automation processing a language L (Casey, 1996; Petersson, 2005; Petersson et al, 2005, 2012; Petersson & Hagoort, 2012; Siegelmann, 1999). If the trajectories associated with the strings of L are learned from input-output data rather than specifically compiled into the network, an ANN with recurrent connections can from a computational point of view be taken to have inferred the automaton that generates L or, equivalently, the grammar of L. Recurrent ANN architectures used for grammatical inference are discrete time, continuous space dynamical systems computationally interpreted as finite state machines.

3.2 Finite state architecture and natural language processing

3.2.1 Generative power

Early in the development of generative grammar Chomsky rejected finite state architecture as an adequate model for natural language (1956, 1957, 1959, 1963; Miller & Chomsky, 1963), and it has been ignored for that purpose ever since; for overviews see (Fitch & Friederici, 2012; Fitch et al, 2012). There are three main reasons for this.

(i) *Weak generative capacity*

Given a finite set of symbols A, the *n*-fold Cartesian product An is the set of all possible sequences of symbols $a \in A$, where *a* is any symbol in A, *n* is a positive integer, and the sequences are strings. In the theory of computation a language is a subset L of An, and one way of selecting L is by specifying an automaton which generates all and only the strings belonging to L. Such language-defining functions have been divided into classes on the basis of what string symbol patterns they are capable of generating; this is the Chomsky Hierarchy (Jäger & Rogers, 2012) in automata theory or, equivalently, of grammars in formal language theory (Hopcroft et al, 2000; Fitch & Friederici, 2012). In generative linguistics the string-set defining capacity of grammars and automata is called weak generative capacity.

Generative grammar assumes that natural languages are infinite string sets because there is no bound on the application of the recursive grammatical rules that generate them (Fitch & Friederici, 2012; Pullum & Scholz, 2010). Center embedding, that is, the string pattern anbn, is of interest because the class of FSAs cannot accept or generate languages which include it. An FSA by definition has a finite number of states.

If it is given enough states to accept anbn for some *n*, and it then encounters an+1bn+1, it will not have enough states to decide on membership of L, and this will always be possible no matter how many states the FSA is given. The class of FSAs cannot, therefore, define languages containing the anbn pattern because *n*, and therefore the length of the pattern, is unbounded. And, as Chomsky observed, natural languages include center-embedding of grammatical constituents, ie, *the cat the dog saw ran = NP2NP1VP1VP2*.

(ii) *Strong generative capacity*

A foundational principle of generative grammar has been that natural language strings have a complex compositional syntax, that is, a structure beyond the strictly temporal or spatial sequentially of speech utterances and text, and more specifically have a recursively compositional phrase structure where the tree diagrams representing the structure must allow simultaneous left and right phrasal nonterminal branching from parent nodes (Fitch et al, 2012; Szabo, 2020).

Such complex phrase structure is fundamental to the explanatory capacity of syntactic theory in linguistics because it allows intuitions and empirical findings about patterns in sequential natural language strings to be expressed in theoretically satisfying generalizations. It is also the foundation on which compositional theories of linguistic meaning are built (Dever, 2006; Szabó, 2020). FSA architecture imposes the same structure on every string, strict sequentiality, and can thereby provide only an impoverished view of natural language syntax and semantics - syntactically, every string is just one word after another, and semantically the meaning of a string is a sequential superposition of the meanings of its constituent words. Chomsky pointed out the impoverished explanatory capacity of FSA architecture in the 1950s, and it has been ignored by generative linguists ever since. Instead, syntactic and semantic theory has been based on automata which are higher than FSAs in the Chomsky Hierarchy and on the corresponding formal grammars because these can support complex phrase structure; finiteness has been accommodated via memory limits on stack depth or tape length, which renders all the automata classes finite state but retains their explanatory capacities.

iii. *Learnability*

Generative linguistics has argued that it is necessary to posit an innate predisposition to learn a natural language, a language acquisition device (LAD), because the linguistic environment in which the child matures is too impoverished to allow the grammar of the language to be inductively inferred *ab initio,* a view that has been controversial; for a critical review of the literature on the LAD and the related notion of Universal Grammar (UG) see (Christiansen & Chater, 2008). Since its publication in 1967, a paper by Gold on the mathematics of language learnability has been invoked in support of the LAD. Gold concluded that only the class of finite cardinality languages were learnable by induction from 'text', that is, solely from examples of strings from the language to be learned, and that, unless one is prepared to regard natural languages as finite, the fact that they are acquired in reality requires something more than 'text', such as prior knowledge about the language or provision of negative as well as positive examples of language

membership. The commitment to the infinity of natural languages in generative linguistics has precluded acceptance of finite cardinality.

S relies relies entirely on inductive learning from 'text'. There is no provision of prior structure to model prior linguistic knowledge, so the indication is that general natural language learning will be impossible for it.

On grounds of weak and strong generative capacity as well as learnability, therefore, S appears to be ruled out a a viable natural natural language model on account of its FSA architecture.

3.3 Resolution

In generative linguistics the foregoing ideas underlie Chomsky's distinction of competence and performance modelling of natural language, where the latter, undertaken by disciplines like psycholinguistics, sociolinguistics, and neurolinguistics, models the human language faculty as implemented in the brain and used by actual speakers in real-time discourse, and the former, undertaken by generative linguistics and philosophy of mind and language, models a simplification of the language faculty from which limitations arising out of physical implementation and usage have been abstracted away. Throughout his career Chomsky has been clear that he is doing competence modelling and strictly excluding performance (Chomsky, 1965, 2000), a position that has been standard in generative linguistics. For present purposes it is important to keep the distinction in mind because objections to FSA architecture assume a competence modelling framework. For discussion of the competence / performance distinction and its relevance to performance models see (Christiansen & Chater, 1992, 1999; Christiansen & MacDonald, 2009; De Vries et al, 2008, 2011, 2012; Petersson & Hagoor,t 2012).

Natural language speakers and their linguistic capabilities are finite, and this, as Fitch & Friederici (2012) observe, renders the mathematical proofs of formal language and automata theory *'technically irrelevant in the real world of finite brains and finite time'* because they depend on the assumption of unboundedness: we have limited lifespans, our brains have limited capacities, and in a lifetime we hear and generate a large but still bounded number of natural language sentences limited in length to a few tens or, at a stretch, a few hundred words. In other words, in a performance modelling context the foregoing objections for FSA architecture do not apply because the assumption of unboundedness on which the disqualification depends is inapplicable. In a performance model the locus of language is the individual human head rather than the abstraction on which competence modelling is based. In fact, performance models can, on the basis of human finiteness, assume not only finite state languages, which can be infinite, but finite-cardinality ones, and every finite-cardinality language can be generated by an FSA. For further discussion of the implications of finiteness for performance modelling see (Petersson, 2005; Petersson et al, 2005; Petersson et al, 2012; Petersson & Hagoort, 2012; Pullum & Scholz, 2010).

- Weak generative capacity: It has been observed often enough that, though it occurs for *n* = 2 in natural language use, center embedding is very rare for *n* = 3, and as far as anyone knows is unattested for *n* > 3; when confronted with artificial strings for *n* > 2, moreover, humans find them difficult and then incomprehensible with increasing *n* (Christiansen, 1992; Christiansen & Chater, 1999; Christiansen & MacDonald, 2009; de Vries et al, 2008, 2011, 2012; Karlsson, 2007).
- Strong generative capacity: From a competence modelling perspective the argument against strictly sequential FSA architecture is unanswerable, but from a performance perspective it is irrelevant. Because automata classes in the Chomsky hierarchy are downward-inclusive, a finite string set does not imply the class membership of the automaton that generated it. For competence modelling a computational class higher than that of strictly-sequential FSAs is appropriate - pushdown automata / context free grammars, say, which are explanatorily more satisfying than an FSA in terms of describing the phrase structure of natural language strings. But how humans choose to understand a natural phenomenon has no necessary implications for how the phenomenon actually works in nature. Just because we find it useful to understand the structure of natural language strings in terms of complex phrase structure, and therefore theorize the generating mechanism as belonging to a computational class higher than the strictly sequential FSA one, does not preclude nature using an FSA generating mechanism. And if a strictly sequential FSA architecture satisfactorily explains intrinsic intentionality in terms of trajectories in a dynamical system state space, there is no reason based on formal language and automata theory to reject it on principle. As Christiansen (1992) puts it, '*recursion in NL is best construed as a descriptive phenomenon rather than a basic processing mechanism*'.
- Learnability: For performance modelling the learnability problem disappears because the individual language learner only ever sees and generates finite string sets, and such sets are learnable from 'text'. Further on this argument see Petersson (2005) and Petersson & Hagoort (2012); for commentary on Gold's results and surveys of work on language learnability since then see Johnson (2004), Lange et al (2008) and Scholz et al (2015).

In his overview of the computational theory of mind, Rescorla (2020) remarks that '*a normal human can entertain a potential infinity of propositions*'. This assertion is simply false, and the case for the validity of S as a viable performance language model is based on its denial.

Conclusion

S is not intended as, and obviously cannot serve as, a general model for implementation of sentence meaning, though the hope is that it can serve as the conceptual basis for one. A full model would have to accommodate a far greater range of syntactic structures than the simple declarative sentences used in the foregoing simulation. It would also have to incorporate the extensive neuroscientific work on the integration of language and object recognition reviewed, for example, 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 (Piccinini 2009, 2016, 2017), but if it is then a CTM model of it must exist. The aim, rather, has been to propose a way of implementing intrinsic linguistic meaning in a physical system, thereby resolving Searle's problem of original intentionality. Decades ago, during the classicism / connectionism debate about the appropriate level of cognitive theorizing (Rescorla, 2020), Fodor and Pylyshyn (1988) proposed that neural models were best understood as implementation-level accounts of cognitive functions. This struck me as reasonable at the time, and it still does; see further (Piccinini & Craver, 2011; Morgan & Piccinini, 2017). On this view, interpretation of S in cognitive terms would make a fully developed version of it a candidate implementation-level model for intrinsic sentence meaning at the cognitive level.

References

Adams, F., Aizawa, K. (2017). Causal theories of mental content, *Stanford Encyclopedia of Philosophy*, (Summer 2017 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/sum2017/entries/content-causal/>.

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

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

Arbib, M. (1987). *Brains, machines, and mathematics*, 2nd ed. Springer.

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

Barsalou, L. (2016a). On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychonomic Bulletin & Review,* 23, 1122–1142.

Barsalou, L. (2016b). Situated conceptualization: Theory and applications. In *Foundations of embodied cognition: Volume 1 Perceptual and emotional* embodiment (pp. 11-37). Psychology Press.

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

Barsalou, L., Dutriaux, L., Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B* 373.

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

Binder J. (2016). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review,* 23, 1096–1108.

Binder, J., Westbury, C., McKiernan, K., Possing, E., Medler, D. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience,* 17, 905–917.

Binder,J., Desai,R., Graves,W., Conant, L. (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*, 33, 130-174. DOI: 10.1080/02643294.2016.1147426

Boeckx, C., Grohmann, K. (Eds.) (2013). *The Cambridge Handbook of biolinguistics*. Cambridge University Press.

Boone, W., Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese,* 193, 1509-34.

Bradie, M., Harms, W. (2020). Evolutionary Epistemology. *Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2020/entries/epistemology-evolutionary/>.

Casey, M. (1996). The dynamics of discrete-time computation with application to recurrent neural networks and finite state machine extraction. *Neural Computation*, 8, 1135–1178.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory,* 2, 113–124.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control,* 2, 137–167.

Chomsky, N. (1963). Formal properties of grammars. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 323–418). John Wiley & Sons.

Chomsky, N. (1965). *Aspects of the theory of syntax.* MIT Press.

Chomsky, N. (2000). The architecture of language, ed. N. Mukherji, B. Patnaik, R. Agnihotri. Oxford University Press.

Christiansen, M. (1992). The (non)necessity of recursion in natural language processing. *Proceedings of the 14th Annual Conference of the Cognitive Science Society.* 665–670.

Christiansen, M., Chater, N. (1999). Towards a connectionist model of recursion in human linguistic performance. *Cognitive Science,* 23, 157-205.

Christiansen, M., Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences,* 31, 489-558.

Christiansen, M., MacDonald, M. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, 59, 126-161.

Churchland, P. (2012). *Plato's camera. How the physical brain captures a landscape of abstract universals*. MIT Press.

Cole, D. (2020). The Chinese Room Argument. *Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), ed. E. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/chinese-room/>.

Conway C., Pisoni D. (2008). Neurocognitive basis of implicit learning of sequential structure and its relation to language processing. *Annals of the New York Academy of Sciences,* 1145, 113-31.

Dever, J. (2006). Compositionality. In E. Lepore & B. Smith (Eds.), *The Oxford Handbook of Philosophy of Language* (pp. 633-666). Oxford University Press.

De Vries, M., Monaghan, P., Knecht, S., Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: the learnability of embedded hierarchical structures. *Cognition,*107, 763–774.

De Vries, M., Christiansen, M., Petersson, K. (2011). Learning Recursion: Multiple Nested and Crossed Dependencies. *Biolinguistics*, 5, 10-35.

De Vries, M., Petersson, K., Geukes, S., Zwitserlood, P., Christiansen, M. (2012). Processing multiple non-adjacent dependencies: evidence from sequence learning. *Philosophical Transactions of the Royal Society B,* 367, 2065-2076.

Downes, S. (2018). Evolutionary Psychology. *Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2020/entries/evolutionary-psychology/>.

Fernandino, L., Binder, J., Desai, R., Pendl, S., Humphries, C., Gross, W., Conant, L., Seidenberg, M. (2016). Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics. *Cerebral Cortex*, 26, 2018–2034.

Fitch, W., Friederici, A. Hagoort, P. (2012). Pattern perception and computational complexity: introduction to the special issue. *Philosophical Transactions of the Royal Society B*, 367, 1925-1932.

Fitch, W., Friederici, A. (2012). Artificial grammar meets formal language theory: an overview. *Philosophical Transactions of the Royal Society B,* 367, 1933-1955.

Fodor J., Pylyshyn Z. (1988). Connectionism and cognitive architecture. *Cognition,* 28, 3–71.

Friederici, A. (2017). *Language in our Brain: The origins of a uniquely human capacity*. MIT Press.

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

Gallistel, C. (2008). Learning and representation. In J. Byrne (Ed.), *Learning and memory: A comprehensive reference* (pp. 227–242). Elsevier.

Gallistel, C. King, A. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. Wiley.

Gärdenfors, P. (2014) *Geometry of meaning. Semantics based on conceptual spaces.* MIT Press.

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, European Journal of Neuroscience, 43, 721-737.

Gazzaniga, M., Ivry, R., Mangun, G. (2019). *Cognitive neuroscience: the biology of the mind.* W. W. Norton.

Geeraerts, D., Cuyckens, H. (2012). *Introducing cognitive linguistics.* Oxford University Press.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology and Philosophy*, 32, 337–355.

Gold, E. (1967). Language Identification in the Limit. *Information and Control,* 10, 447–474.

Goodfellow, I., Bengio, Y. (2017). *Deep Learning.* MIT Press.

Haykin, S. (2008). *Neural Networks: A Comprehensive Foundation.* Prentice Hall.

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation,* 9, 1735-1780.

Hopcroft, J., Motwani, R., Ullman, J. (2000). *Introduction to automata theory, languages, and computation. Reading.* Addison-Wesley.

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

Jackendoff, R. (2002). Foundations of Language: Brain, Meaning, Grammar, Evolution. Oxford University Press.

Jackendoff, R. (2012). A User's Guide to Thought and Meaning. Oxford University Press.

Jacob, P. (2019). Intentionality. *Stanford Encyclopedia of Philosophy*, (Spring 2019 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2019/entries/intentionality/>.

Jäger, G., Rogers, J. (2012). Formal language theory: refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B*, 367, 1956–1970.

Johnson, K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, 70, 571–592.

Karlsson, F. (2007). Constraints on multiple initial embedding of clauses. *International Journal of Corpus Linguistics,* 12, 107-118.

Kemmerer, D. (2015). *Cognitive Neuroscience of Language*. Psychology Press.

Lange, S., Zeugmann, T., Zilles, S. (2008). Learning indexed families of recursive languages from positive data: a survey. *Theoretical Computer Science,* 397, 194-232.

Lau, J., Deutsch, M. (2014). Externalism About Mental Content. *Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/fall2019/entries/content-externalism/>.

Macleod, C. (2016). John Stuart Mill. *Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/sum2020/entries/mill/>.

Matheson, H., Barsalou, L. (2018). Embodiment and grounding in cognitive neuroscience. In J. Wixted, (Ed.) *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (4th ed.).* Wiley.

McLaughlin, B., Bennett, K. (2018). Supervenience. *Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/win2018/entries/supervenience/>.

Miller, G., Chomsky, N. (1963). Finitary models of language users. In D. Luce, R. Bush, E. Galanter, (Eds.), *Handbook of Mathematical Psychology 2* (419-91). John Wiley & Sons.

Millikan R. (2004). *Varieties of Meaning*. MIT Press.

Moisl, H. (2020). Intrinsic Intentionality and Linguistic Meaning: An Historical Outline. In E. Kelih, R. Köhler (eds.) *Words and Numbers. In Memory of Peter Grzybek 1957-2019 (*148-166*).* RAM-Verlag.

Morgan, A., Piccinini, G. (2017). Towards a cognitive neuroscience of intentionality. *Minds and Machines,* 28, 119-139.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT Press.

O'Connor, T. (2020). Emergent Properties. *Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), ed. E. Zalta, forthcoming URL = <https://plato.stanford.edu/archives/fall2020/entries/properties-emergent/>.

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

Papineau, D. (2020). Naturalism. *Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/sum2020/entries/naturalism/>.

Patton, L. (2018). Hermann von Helmholtz. *Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/win2018/entries/hermann-helmholtz/>.

Petersson, K. (2005). On the relevance of the neurobiological analogue of the finite-state architecture. *Neurocomputing,* 65-66, 825-832.

Petersson, K., Grenholm, P., Forkstam, C. (2005). Artificial grammar learning and neural networks. *Proceeding of the Cognitive Science Society*, 2005, 1726–1731.

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

Petersson, K., Hagoort, P. (2012). The neurobiology of syntax: beyond string sets. *Philosophical Transactions of the Royal Society B,* 367, 1971-1983.

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

Piccinini, G. (2015). *Physical computation. A mechanistic account*. Oxford University Press.

Piccinini, G. (2016). The computational theory of cognition. In V. Müller (Ed.) *Fundamental issues in artificial intelligence*. Springer.

Piccinini, G. (2017). Computation in Physical Systems, *Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), ed. E. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/computation-physicalsystems/>.

Piccinini, G., Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science,* 37, 453-488.

Piccinini, G., Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese,* 183, 283–311.

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

Piccinini, G., Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics,* 37, 1-38.

Piccinini, G., Shagrir, O. (2014). Foundations of computational neuroscience. *Current Opinion in Neurobiology,* 25, 25-30.

Plebe, A., de la Cruz, V. (2016). *Neurosemantics. Neural processes and the construction of linguistic meaning*. Springer.

Pojman, P. (2019). Ernst Mach. *Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2019/entries/ernst-mach/>.

Pullum, G., Scholz, B. (2010). Recursion and the infinitude claim. In H. van der Hulst (Ed.), *Recursion and Human Language* (pp.111-138). De Gruyter Mouton.

Pulvermüller, F. (2012). Meaning and the brain: The neurosemantics of referential, interactive, and combinatorial knowledge. *Journal of Neurolinguistics,* 25, 423-459.

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

Ramsey, W. (2019). Eliminative Materialism. *Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/sum2020/entries/materialism-eliminative/>.

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

Rescorla, M. (2020). The computational theory of mind. *Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2020/entries/computational-mind/>.

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

Ryder D. (2004). SINBAD neurosemantics: a theory of mental representation. *Mind and Language,* 19, 211–240.

Rysiew, P. (2020). Naturalism in Epistemology. *Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), ed. E. Zalta, forthcoming URL = <https://plato.stanford.edu/archives/fall2020/entries/epistemology-naturalized/>.

Scholz, B., Pelletier, F., Pullum, G. (2015). Philosophy of Linguistics. *Stanford Encyclopedia of Philosophy*, (Winter 2016 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/win2016/entries/linguistics/>.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3, 417–57.

Shagrir, O. (2018). The Brain as an input–output model of the world. *Minds and Machines,* 28, 53-75.

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

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

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Siegelmann, H., Sontag, E. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences,* 50, 132-50.

Siegelmann, H. (1999). *Neural networks and analog computation: beyond the Turing limit*. Birkhäuser.

Sipser, M. (2012). *Introduction to the theory of computation*, International edition. Thomson.

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

Speaks, J. (2021). *Theories of meaning*. *Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2021/entries/meaning/>.

Stoljar, D. (2015). Physicalism. *Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/win2017/entries/physicalism/>.

Szabó, Z. (2020). Compositionality. *Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/fall2020/entries/compositionality/>.

Tangirala, A. (2014). *Principles of System Identification: Theory and Practice.* CRC Press.

Thomson, E., Piccinini, G. (2018). Neural representations observed. *Minds and Machines,* 28, 191-235.

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.

van Riel, R., van Gulick, R. (2019). Scientific Reduction. *Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2019/entries/scientific-reduction/>.

Wilson, R., Foglia, L. (2015). Embodied Cognition. *Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2017/entries/embodied-cognition/>.

Declarations

Funding: Self-funded

Conflicts of interest/Competing interests: not applicable

Availability of data and material: To be posted on Trolling upon acceptance

Code availability: To be posted on Github upon acceptance

Ethics approval: not applicable

Consent for publication: not applicable

Figures

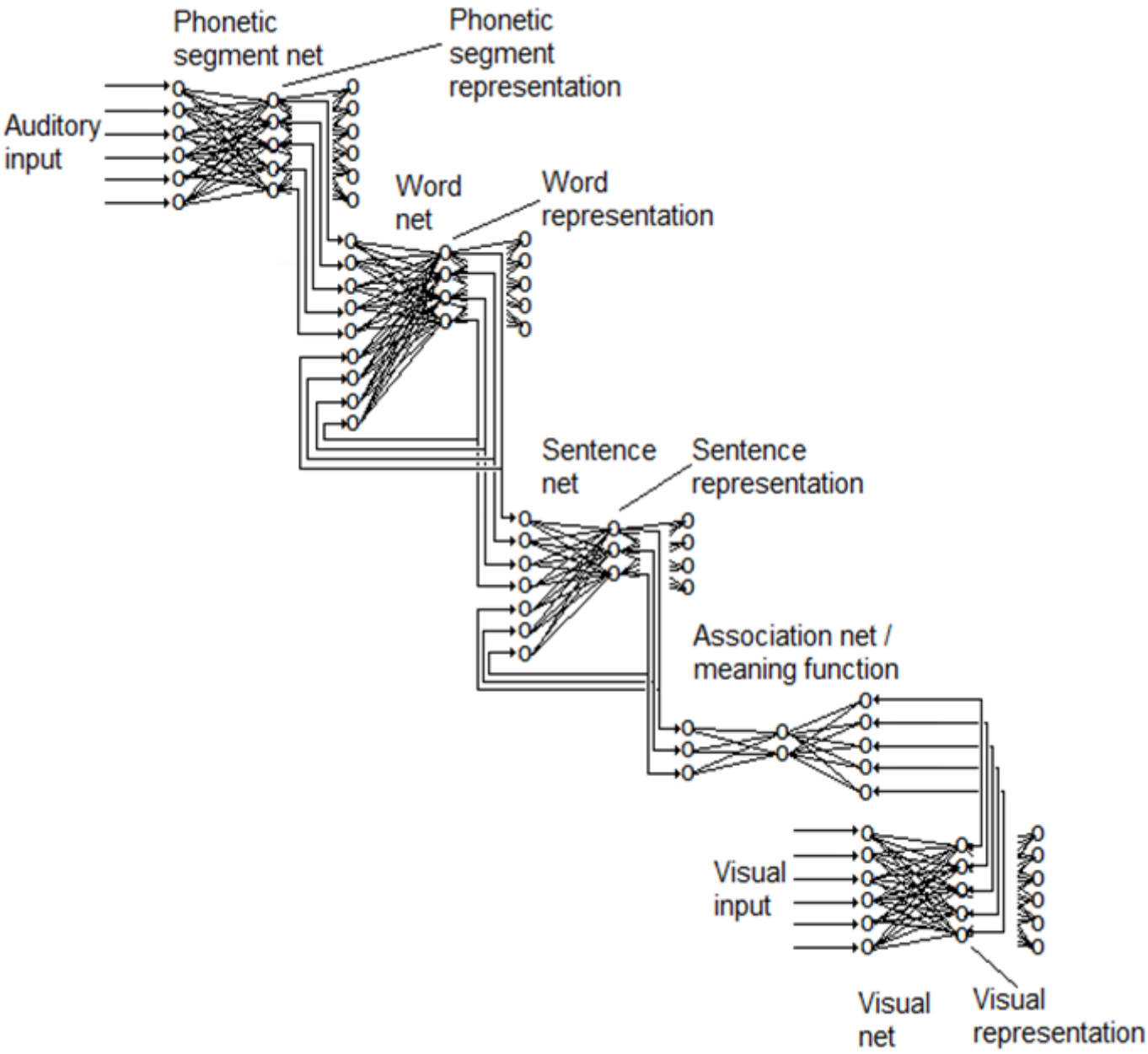


Figure 1

Structure of S

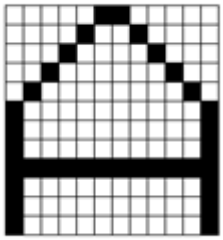


Figure 2

A letter bitmap

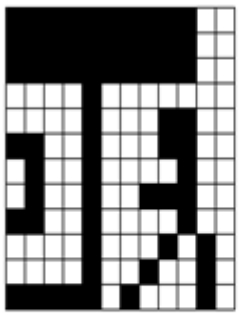
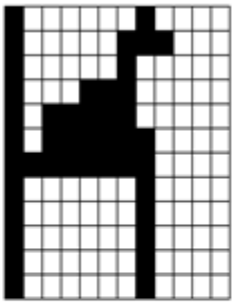
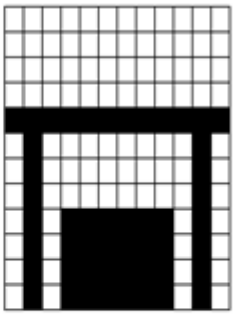
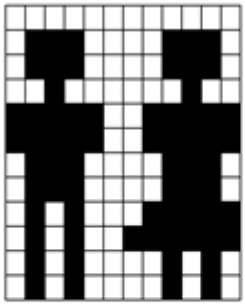
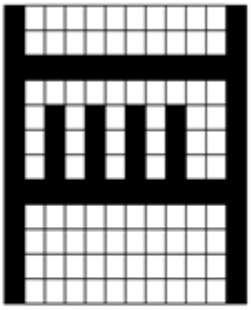
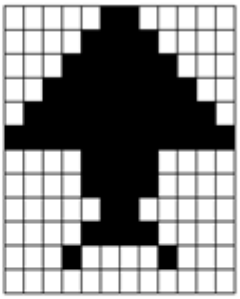
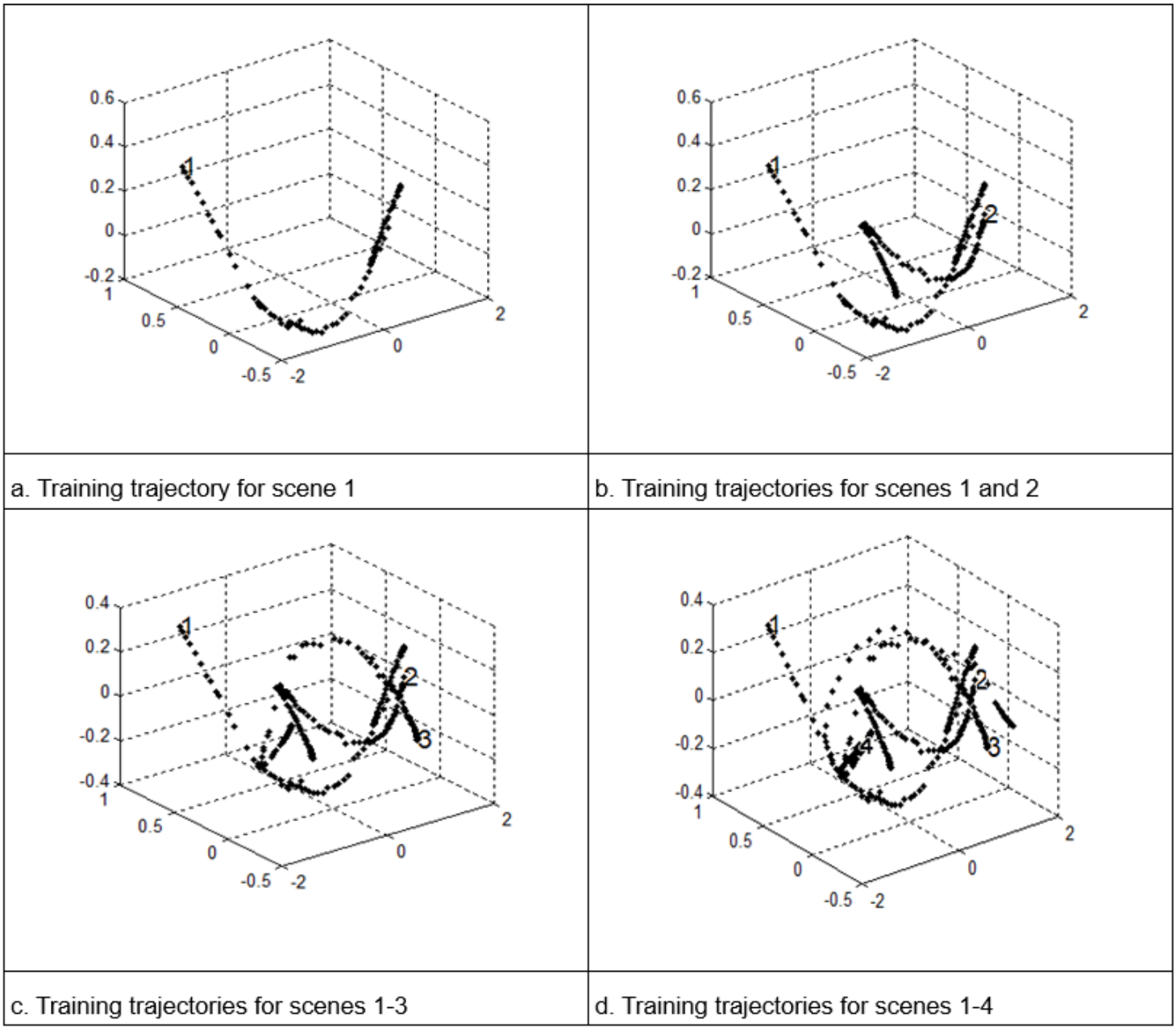
		
1. the man goes into the house	2. the dog sits on the chair	3. the chest is under the table
		
4. the boy stands beside the girl	5. the books are on the shelf	6.the jet plane flies over head

Figure 3

Simulation training set



a. Training trajectory for scene 1

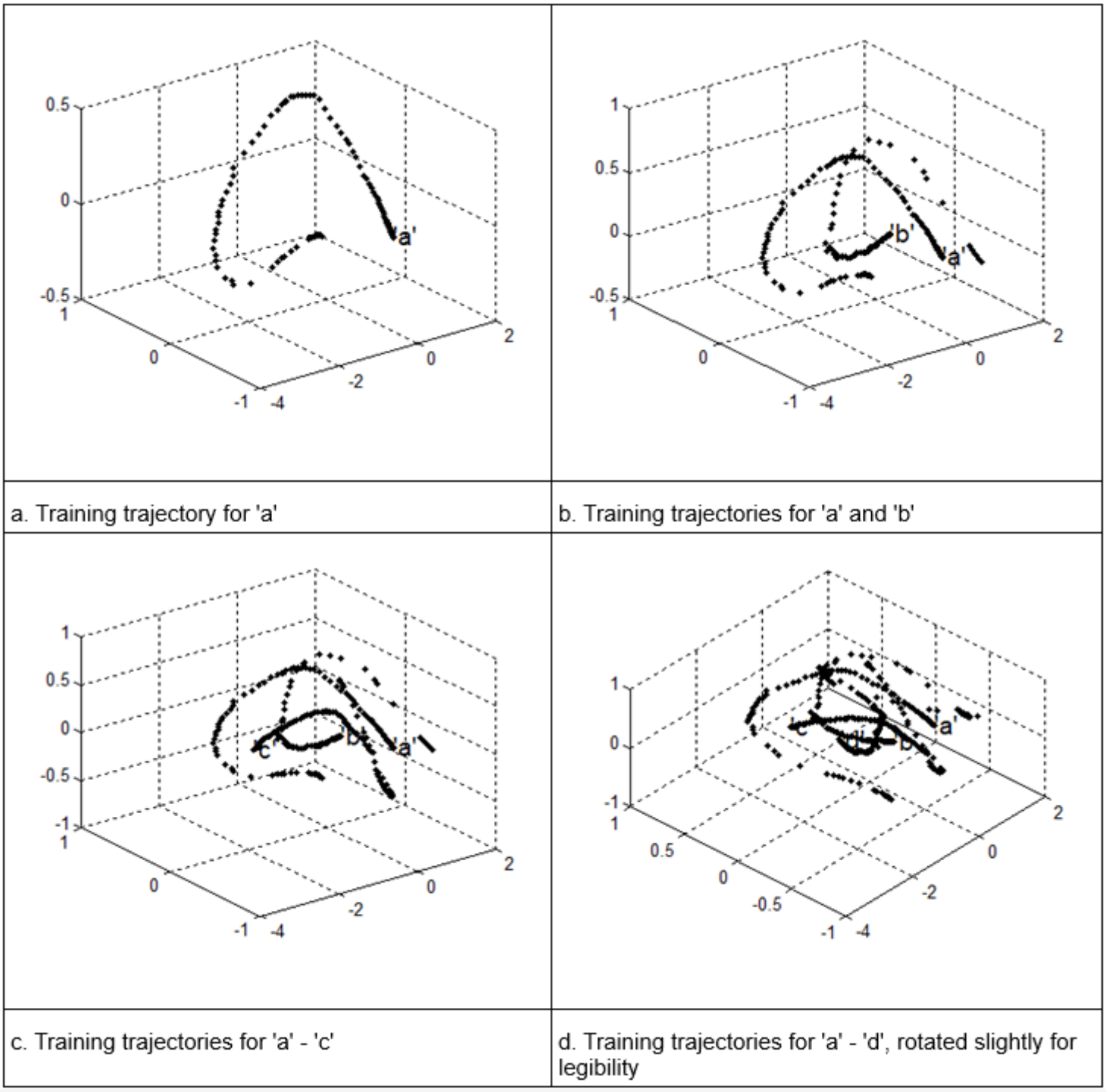
b. Training trajectories for scenes 1 and 2

c. Training trajectories for scenes 1-3

d. Training trajectories for scenes 1-4

Figure 4

Visual subnet training trajectories



a. Training trajectory for 'a'

b. Training trajectories for 'a' and 'b'

c. Training trajectories for 'a' - 'c'

d. Training trajectories for 'a' - 'd', rotated slightly for legibility

Figure 5

Letter training trajectories

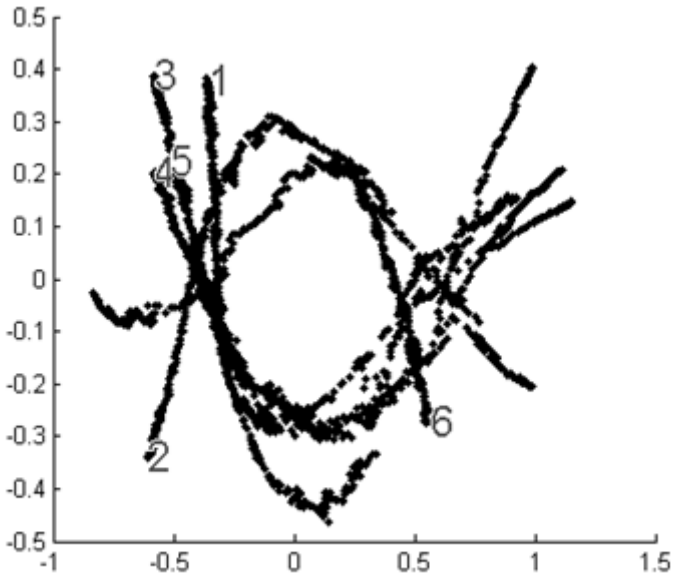


Figure 6

Association subnet hidden layer training trajectories

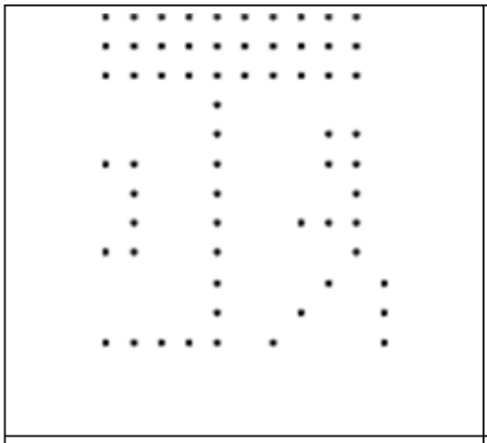
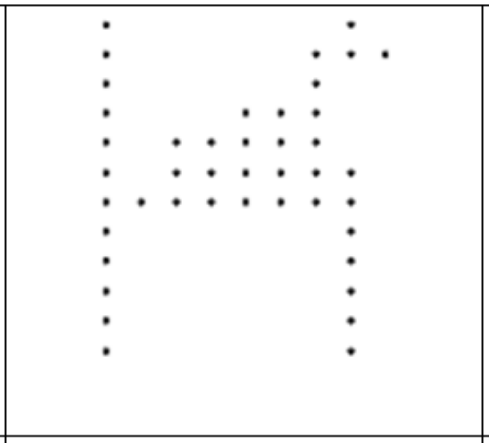
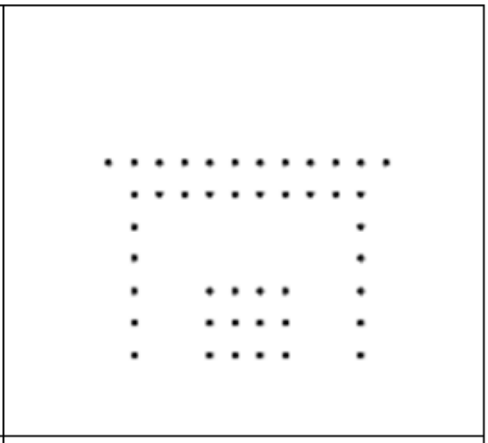
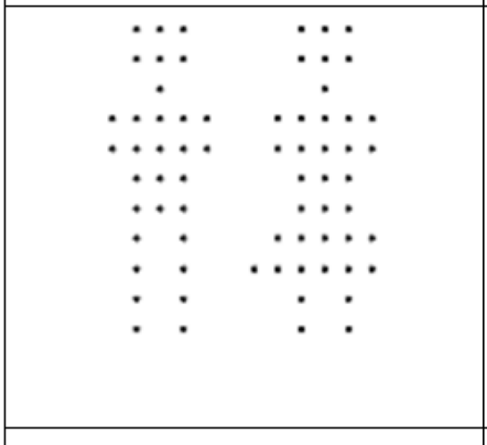
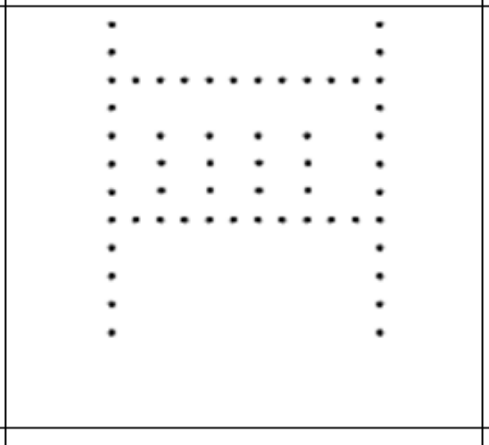
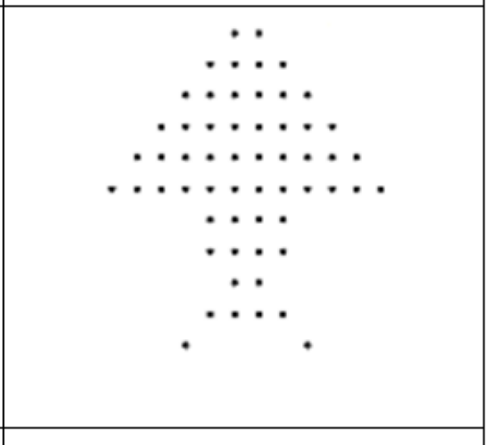
		
1. the man goes into the house	2. the dog sits on the chair	3. the chest is under the table
		
4. the boy stands beside the girl	5. the books are on the shelf	6. the jet plane flies over head

Figure 7

Output from S after training

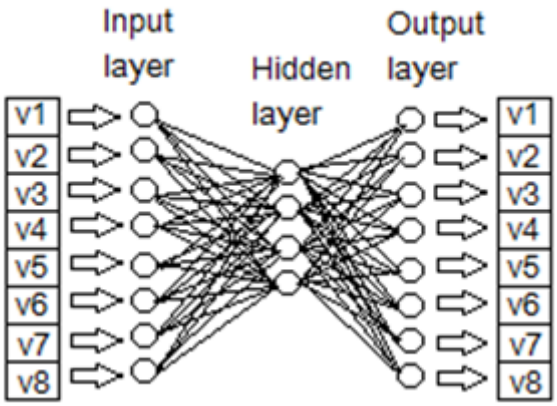


Figure 8

An autoassociative MLP

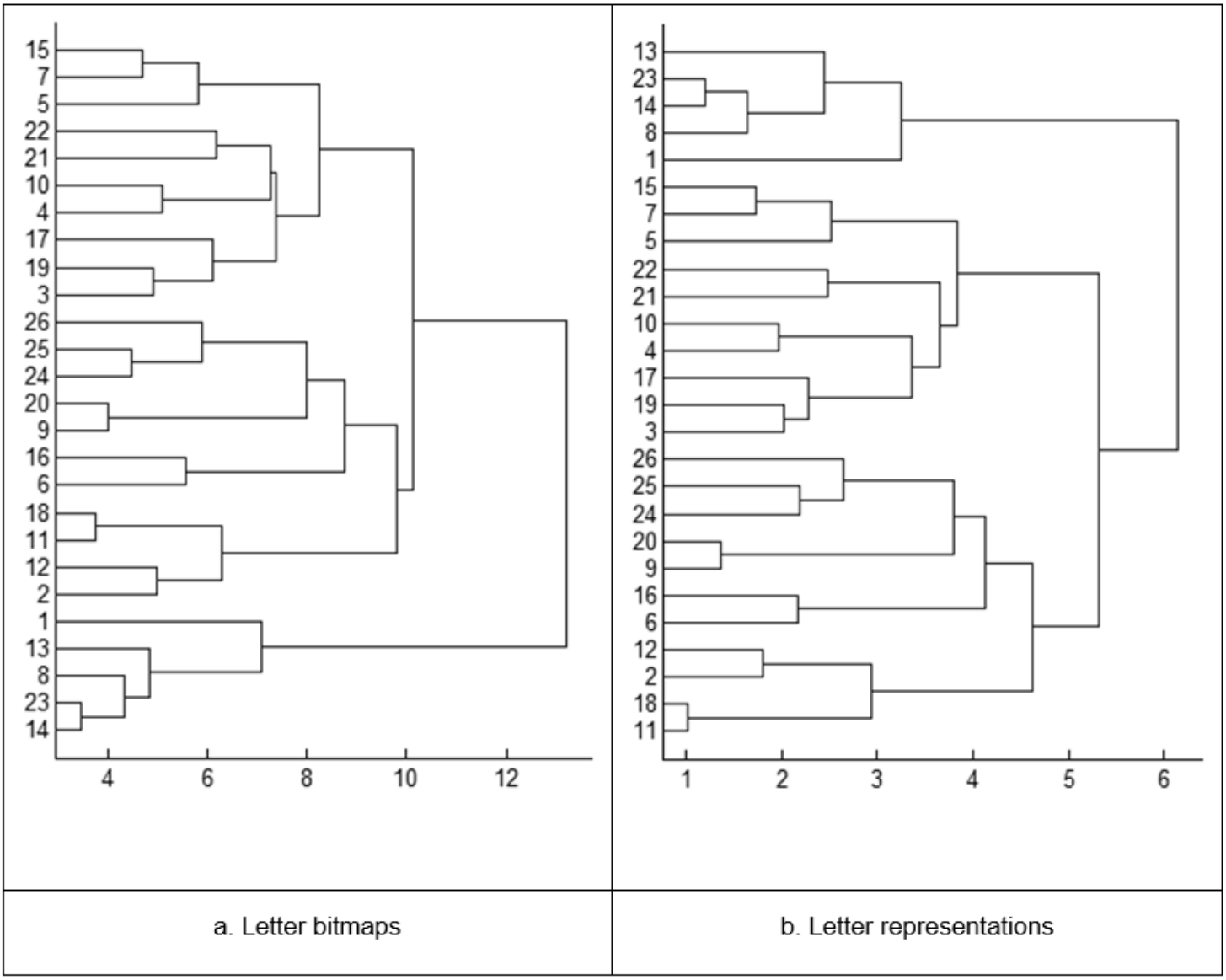


Figure 9

Cluster trees for letter bitmaps and their hidden layer representations

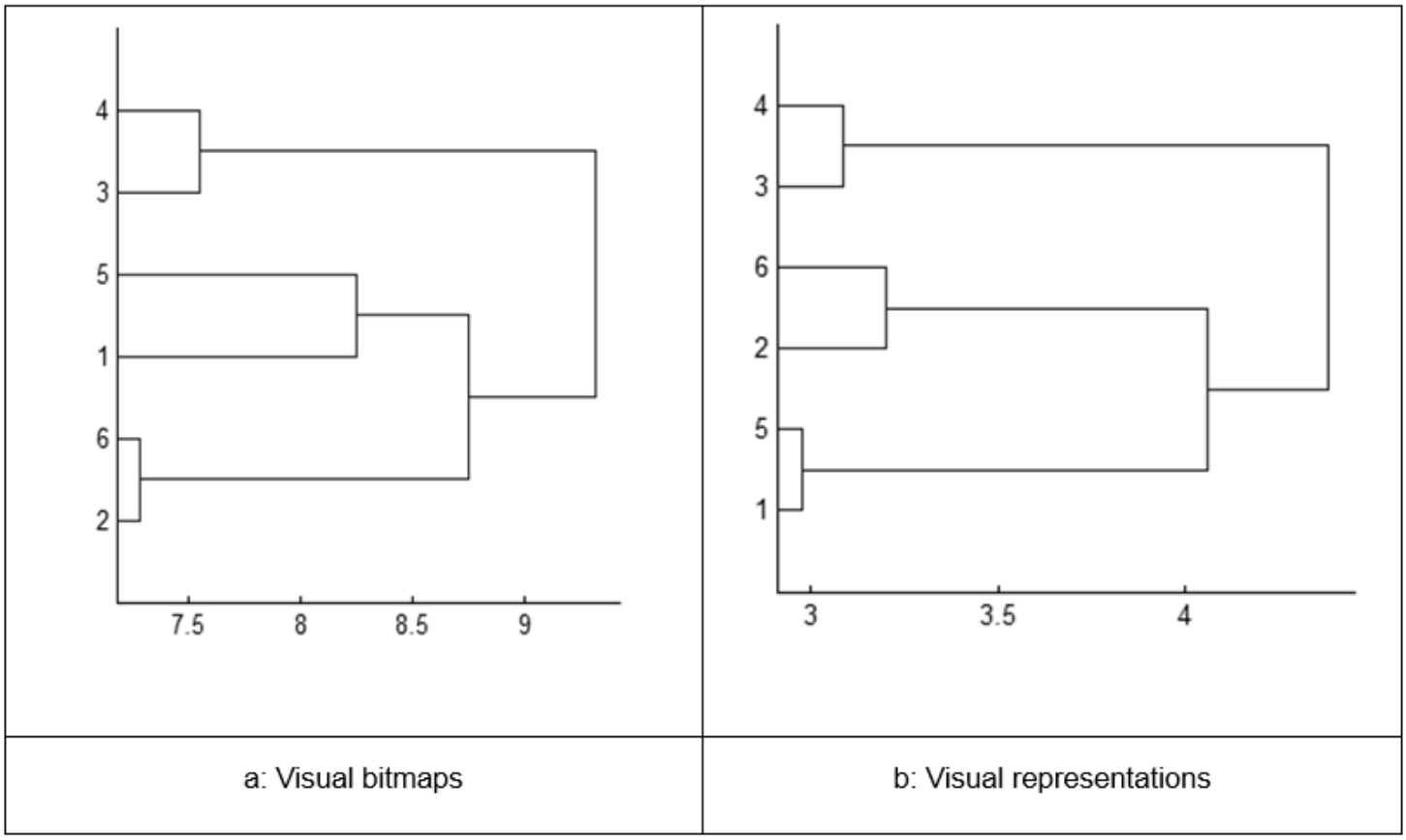


Figure 10

Cluster trees for visual bitmaps and their hidden representations. The numbers refer to the bitmap images in Figure 3

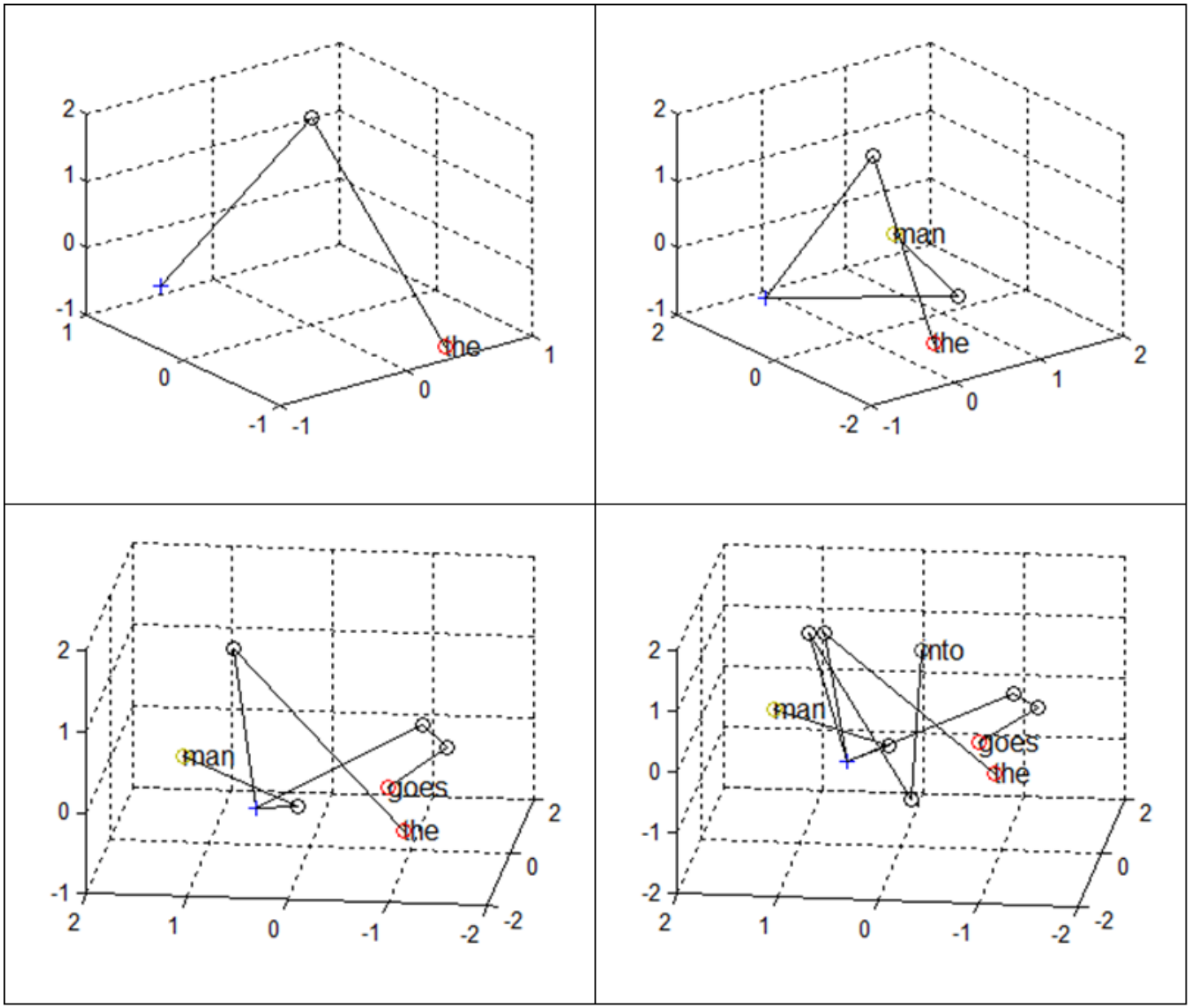


Figure 11

Word subnet hidden layer trajectories

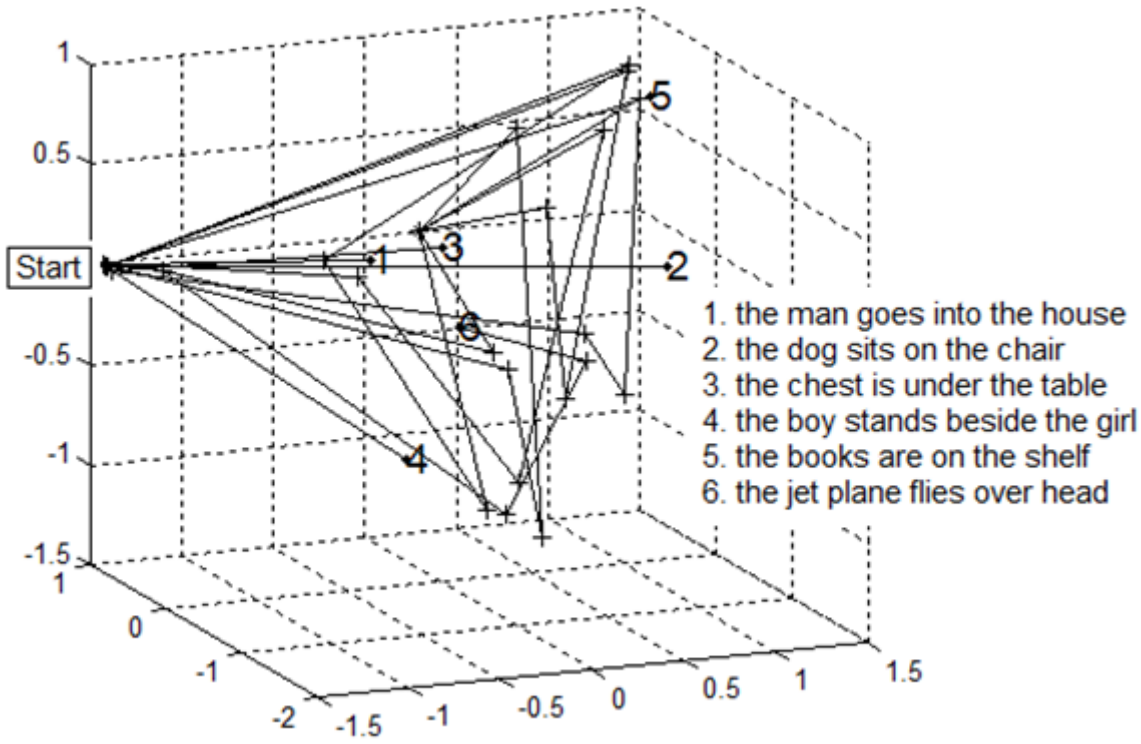


Figure 12

Sentence subnet hidden layer trajectories

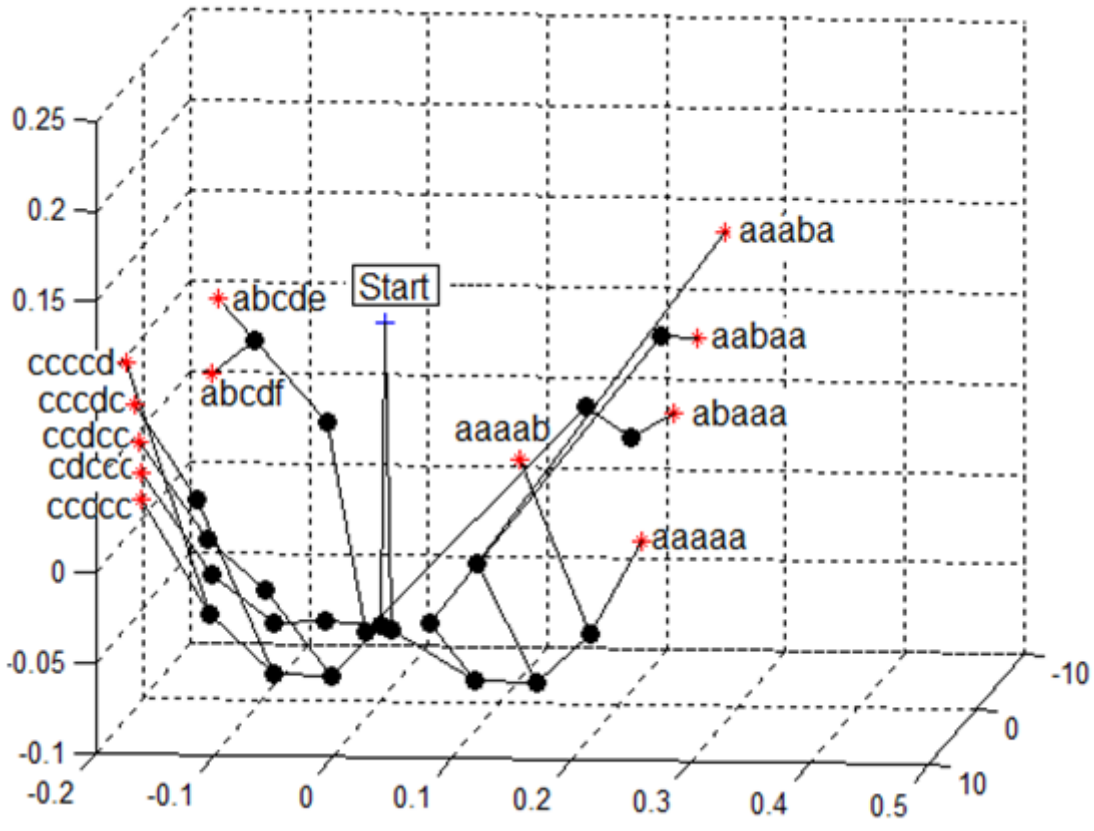


Figure 13

Alphabetic string trajectories

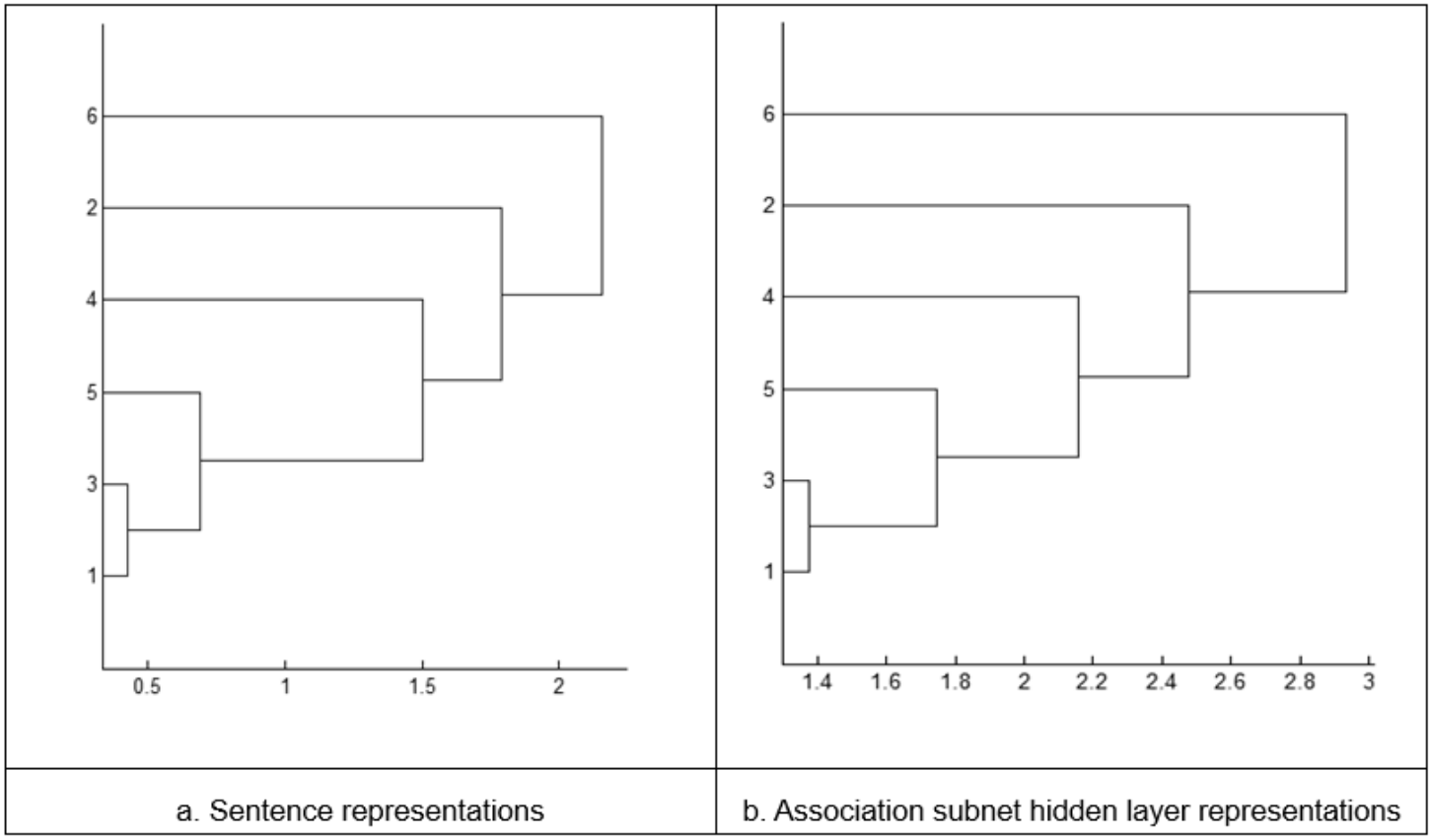


Figure 14

Cluster trees for sentence representations and their hidden layer representations in the association subnet