Connection Science

ISSN: 0954-0091 (Print) 1360-0494 (Online) Journal homepage: http://www.tandfonline.com/loi/ccos20

The importance of situation-specific encodings: analysis of a simple connectionist model of letter transposition effects

Shin-Yi Fang, Garrett Smith & Whitney Tabor

To cite this article: Shin-Yi Fang, Garrett Smith & Whitney Tabor (2017): The importance of situation-specific encodings: analysis of a simple connectionist model of letter transposition effects, Connection Science, DOI: 10.1080/09540091.2016.1272097

To link to this article: http://dx.doi.org/10.1080/09540091.2016.1272097

Published online: 16 Jan 2017.

CONNECTION SCIENCE, 2017
http://dx.doi.org/10.1080/09540091.2016.1272097

The importance of situation-specific encodings: analysis of a simple connectionist model of letter transposition effects

Shin-Yi Fang[a,b†], Garrett Smith[a] and Whitney Tabor[a,b]

[a]Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA; [b]Haskins Laboratories, New Haven, CT, USA

ABSTRACT

This paper analyses a three-layer connectionist network that solves a translation-invariance problem, offering a novel explanation for transposed letter effects in word reading. Analysis of the hidden unit encodings provides insight into two central issues in cognitive science: (1) What is the novelty of claims of "modality-specific" encodings? and (2) How can a learning system establish a complex internal structure needed to solve a problem? Although these topics (embodied cognition and learnability) are often treated separately, we find a close relationship between them: modality-specific features help the network discover an abstract encoding by causing it to break the initial symmetries of the hidden units in an effective way. While this neural model is extremely simple compared to the human brain, our results suggest that neural networks need not be black boxes and that carefully examining their encoding behaviours may reveal how they differ from classical ideas about the mind-world relationship.

ARTICLE HISTORY

Received 6 April 2016
Accepted 11 December 2016

KEYWORDS

Connectionism; learning; modality-specific encodings; symmetry breaking; letter transposition

1. Introduction

Over the past several decades, various theoretical perspectives in cognitive science have urged the field to move away from the symbolic computational approach which first galvanised it in the mid-twentieth century: connectionism (Churchland & Sejnowski, 1992; McClelland, Rumelhart, & the PDP research group, 1986; Rumelhart, McClelland, & the PDP research group, 1986) rejects localist and discrete representations and emphasises an important role for feedback effects; dynamical field theory (Johnson, Spencer, & Schöner, 2008; Thelen & Smith, 1994) argues for the relevance of dynamical systems phenomena like attractors and self-organisation; ecological psychology (Michaels & Carello, 1981; Prindle, Carello, & Turvey, 1980) advocates "direct" (unmediated by inferential processes) perception and also promotes dynamical systems theory (Kelso, 1995); embodied cognition (Barsalou, 2003; Barsalou, Simmons, Barbey, & Wilson, 2003; Glenberg & Kaschak, 2002) argues for perceptual simulation in place of symbolic abstraction. These perspectives have a number of commonalities. They tend to emphasise interactive, feedback models,

CONTACT Shin-Yi Fang ✉ shin-yi.fang@gmail.com

†Present address: Department of Psychology, Pennsylvania State University, 454 Moore Building, University Park, PA 16802, USA

often related to neural structure. They emphasise the groundedness of abstract mental concepts in concrete physical properties of the world. All four approaches place particular importance on the context-sensitivity of an organism's behaviour.

However, it is not yet clear if the new approaches constitute a significant divergence from traditional assumptions. For example, work in embodied cognition finds that when people think of concepts like "lawn", "cranberry", "wrench", etc., they simulate specific instantiations of these entities, influenced by their context of occurrence (e.g. "lawn" evokes roots in the context "rolled-up lawn" but it evokes blades of grass in the context "front lawn", Wu & Barsalou, 2009). Relatedly, other studies find that sensory (Martin, 2007; Martin & Chao, 2001; Simmons, Hamann, Harenski, Hu, & Barsalou, 2008) and motor-related (Beauchamp, 2005; Beauchamp, Lee, Haxby, & Martin, 2002) areas of the cerebral cortex are involved in the neural response to concepts even when perception and action are not obviously involved. But even though behavioural and neural studies show that people are activating specific properties when they process words or images with abstract import, it is possible that these context-sensitive activations do not add anything particularly important to the claims of the classical view. Perhaps abstract symbolic representations are doing the main work of organising our thought and allowing us to make useful inferences; for example, the situation-specific encoding may be a kind of peripheral resonance stemming from bi-directional connections between the seat of abstraction in the neural core and the sensory layers closer to the periphery (the bidirectional connections being needed, in any case, to permit both bottom-up and top-down information flow, for example, for detection and prediction) (Mahon & Caramazza, 2008). It might be that the system would exhibit the same abilities, make the same decisions, and occupy the same ecological niche even if this peripheral resonance were not present.

Another argument against the importance of situation-specific encodings is that veridical simulation of the world is of little relevance to perceptual theory. Positing simulation of the world[1] simply transfers the problem of understanding how the brain interprets the world around it to the problem of understanding how the brain interprets its mental simulation of the world around it. Therefore, according to this argument, while these approaches may have demonstrated the existence of context-sensitivity as a low-level neural and behavioural phenomenon, this finding has little relevance to perceptual theory.

Here, we take a close look at a connectionist model which develops (via a learning process) context-sensitive encodings of orthographic words. We created a model that is sufficiently complex to handle a theoretically interesting case but also sufficiently simple that we can analyse it carefully. By showing how context-specific encodings can play a central role in the functionality of an artificial neural model, our findings suggest, against the arguments just enumerated, that context-specific encodings may be centrally relevant to perception; they also suggest that a more precise understanding of the nature of this context-sensitivity is needed. In particular, we find (1) that although the model forms encodings that mirror seemingly irrelevant physical details, the mirroring is not veridical; it is warped with respect to the properties of the world to which we standardly apply the label "physical". Thus it may not be appropriate to describe what it does as "simulation" or "representation" of the world. We also find (2) that although we trained the model to perform an abstraction (assigning constant phonology to written words when they appear in different spatial positions), the model learned successfully only if there was a physical asymmetry in the data set (some category members were more typical than others); this asymmetry was not important as far as the abstract behaviour was concerned (there was no difference in behaviour required of the atypical category members), but the asymmetry played a crucial role in allowing the learning to succeed. This finding suggests an interdependence of symbolic and subsymbolic structure that argues against classical conceptions of representational abstraction (see Fodor & McLaughlin, 1990; Fodor & Pylyshyn, 1988; cf. Smolensky, 1988, 1991). If the model resembles real people in relevant regards, then these results suggest that the new approaches are onto something.

2. Case study: transposed letter effects

A case of particular interest in word recognition is transposed-letter (TL) effects. Chambers (1979) found that in a lexical decision task, participants were slower to accept words like SLAT that were related to other words by the transposition of two letters (here, SALT) than they were to accept frequency-matched control words (e.g. HUMP). Chambers also found that participants were slower to reject TL non-words (e.g. STROE from STORE) than they were to reject controls derived from frequency-matched words (e.g. CHROB). In a priming paradigm, Forster, Davis, Schoknecht, and Carter (1987) found that for the same target (e.g. INVOLVED), TL non-words (e.g. INVOVLED) produced stronger priming effect than replaced letter (RL) non-words (e.g. INVORVED; see also Schoonbaert & Grainger, 2004). Acha and Perea (2008) found inhibitory effects for words with higher frequency TL neighbours compared with words with no TL neighbours. A general interpretation consistent with these findings is that the mental encodings of the TL stimuli are closer to the encodings of the corresponding base words than are the encodings of the RL stimuli.

The existence of the TL/RL contrast is challenging for traditional slot-based coding schemes (e.g. McClelland & Rumelhart, 1981). Because these schemes separately encode the properties of each letter in each slot, a letter in position 1 is no more similar to the same letter in position 2 than to a different letter in position 1, so these models fail to predict the observed TL/RL contrast. Therefore, a number of researchers (e.g. Davis & Bowers, 2004, 2006; Gomez, Ratcliff, & Perea, 2008; Grainger & Van Heuven, 2003; Whitney, 2001) have proposed distributed coding schemes in which letters are not strictly associated with a single slot, but have ties to nearby slots as well.

Instead of hand-wiring such a distributed code, Rueckl, Fang, Begosh, Rimzhim, and Tobin (2008) created a learning model that was trained to recognise words in multiple horizontally displaced positions. An advantage of considering a learning model rather than a hand-wired model is that there may be subtle properties of the encoding system which stem from the geometry of the task and which we are not likely to be able to formulate intuitively. The learning model provides a quantitatively explicit hypothesis about how the geometric relationships among members of the training set give rise to an encoding geometry. The learning model of Rueckl et al. also ties these TL/RL results to a wide range of other phenomena in word recognition that appear to stem from the interaction of the learning mechanism with properties of the distribution of forms in the language (e.g. non-word naming, frequency by regularity interaction; Davis & Bowers, 2004; Plaut, McClelland, Seidenberg, & Patterson, 1996; varieties of dyslexia, Harm & Seidenberg, 2004).

Here, we review Rueckl et al.'s relatively large simulation of the processing of a sample of English words, noting that it predicts the observed TL/RL contrast. We then describe a highly simplified model that predicts the TL/RL contrast in a similar way to the Rueckl et al. model.

The advantage of studying the simplified model is that it is complex enough to reveal interesting subtleties of the learning dynamics and encoding geometry but low-dimensional enough that we can analyse it in some depth.

2.1. A "large" connectionist model of TL effects

Rueckl et al. (2008) trained a feedforward connectionist model to identify the phonology of 2998 monosyllabic English words, presenting the orthographic form of each word in randomly varying positions. The output layer in their model employed the encoding scheme of Plaut et al. (1996): a Consonant–Vowel–Consonant template with a total 61 phoneme units (23 for the first consonant; 14 for the vowel; 24 for the second consonant). The phoneme units encoded phonemic features like /a/ in POT and /e/ in BED. The input layer used slots for letter positions. Each slot was made up of 26 units corresponding to the 26 letters of the English alphabet (a localist encoding at the letter level). Twelve slots were used to represent 12 horizontally-arrayed letter positions. The presence of a given letter was indicated by setting the appropriate unit to the value of 1 and all others in the slot to 0.

The input layer was fully feedforward-connected to a hidden layer with 200 units, which, in turn, was fully feedforward-connected to the output layer. A unit's net input, net$_{j,}$ was calculated by Equation (1), where a_i represents the activation of unit i, w_{ij} is the weight from unit j to unit i and b_i is the bias of unit i. The activation (a_i) of each unit i was the standard logistic function of the unit's net input net$_i$ as indicated by Equations (1) and (2).

$$\text{net}_i = \sum_j a_i w_j + b_i, \tag{1}$$

$$a_i = \frac{1}{1 + \exp(-\text{net}_i)}. \tag{2}$$

2.1.1. Training procedure

Rueckl et al. used the back-propagation algorithm (Bryson & Ho, 1975; Rumelhart, Hinton, & Williams, 1986) with a cross-entropy cost function (Hinton, 1989) to train this network. On each epoch, the entire training corpus was fed into the model. At the end of an epoch, the weight changes were administered in proportion to a combination of the accumulated error derivative and previous weight changes. The momentum parameter was set to 0.0 initially and then to 0.9 after the first 10 epochs of training (Jacobs, 1988). The weight change was scaled by a global learning rate of 0.001. To discourage overtraining, the reachable targets of 0.1 and 0.9 were used instead of 0 and 1. Weight decay was also added in the training by multiplying the sum of the squares of each weight by a constant of 0.9999 (Hinton, 1989). The small initial weights were assigned random values uniformly distributed between −0.1 and 0.1.

Rueckl et al. found that after training, the average hidden layer (Euclidean) distance between base words and TL non-words was significantly shorter than between base words and RL non-words. To make a more explicit model of the TL priming effect, simulated reaction times were calculated by cascading output unit activations, a method of translating the hidden-output map of the fully trained network into a set of differential equations which exhibit different convergence times for different inputs (McClelland, 1979). TL primes produced significantly shorter reaction times than RL primes. The coincidence of smaller distances between hidden representations and shorter reaction times in the priming paradigm suggests that the source of the reaction time difference is the deployment of the hidden representations. In the following, we make the simplifying assumption that there is a causal relationship between hidden unit geometry and reaction times in the priming paradigm. We therefore focus our effort on understanding the causes (in the interaction of the training data with the learning process) of the hidden unit geometry.

Next, we describe a simple analogy of Rueckl et al.'s model.

2.2. The basic model

2.2.1. Encoding and architecture

We constructed a scaled-down version of Rueckl et al.'s model using an alphabet of only two letters (here called "B" and "D") and "words" only two-letters long. All 4 possible words formed from the alphabet were presented to a scaled-down network (with one unit turned on to uniquely encode each letter in a particular position). The words were presented in each of 9 horizontal positions (Figure 1). In other words, the model had a slot-based "visual field" with a diameter of 10 slots. There were two units in each slot making a total of 20 units in the input layer. A given word, like BB could appear in any one of 9 positions across these slots (e.g. BB-------- = Position 1, -BB------- = Position 2, ... , --------BB = Position 9). The output "phonological code" was identical to the input code except that there was no variation in spatial position. Thus there were just 4 output units: "B" in first position, "D" in first position, "B" in second position, and "D" in second position. Note that neither the input coding nor the output coding provided a simple cue to what we, as readers of written language, think of as physical position. For example, in the input space, the encodings of stimuli BB3 through BB9, which are arrayed at equal intervals along a horizontal line in physical space, were all equally distant from BB1 because these words all had no common input features with BB1. On the other hand, BB1 and BB2 share a feature. In this sense, information indicating the physical proximity of nearby words is present, but this information does not specify the geometric relationships among words offset from one another by more than one slot. The output encoding was position-invariant so it trivially provided no cue to position.

The network had 10 hidden units. Except for the small number of letters, the slightly reduced number of slots (10 instead of 12), and the localist coding output phonemes, this "Basic Model" had the same structure as the model of Rueckl et al.

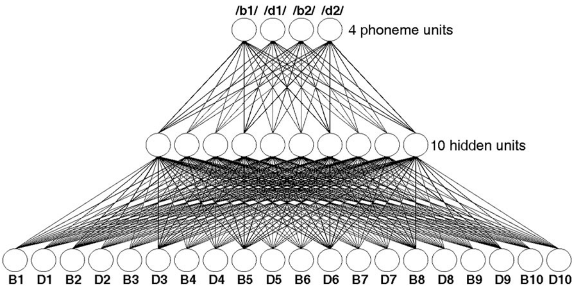
/b1/ /d1/ /b2/ /d2/
4 phoneme units
10 hidden units
B1 D1 B2 D2 B3 D3 B4 D4 B5 D5 B6 D6 B7 D7 B8 D8 B9 D9 B10 D10

Figure 1. Network architecture. The input layer (bottom) represents 10 input slots with letter B and D. There are 10 hidden units in the middle. The output layer (top) represents "phonological code" with no variation in spatial position.

2.2.2. *Training process*

Like Rueckl et al. (2008) we employed the cross-entropy error function and trained the model using backpropagation. During training, each of the 4 words was randomly presented in one of the 9 positions in each epoch. The model was trained in 10 separate runs, each starting with different random initial weights.

2.2.3. *Performance assessment*

The model's performance was first assessed via pronunciation accuracy. A response was counted as accurate when the set of active output units (i.e. activation > 0.5) matched the active target units. After 100,000 epochs of training, all 10 runs produced 100% correct pronunciation for all 4 words in all positions. The average training epoch at which 100% correct behaviour was reached was 44880.

In the simple 4-word vocabulary (BB, BD, DB, DD) we took BD and DB as an analog of a TL pair and BB and DD as an analog of a RL pair. Following Rueckl et al., we defined the similarity of each pair as the average Euclidean distance between corresponding position specific tokens of the pair in all positions (e.g. $\mathit{Similarity}(\mathrm{BB},\ \mathrm{DD}) = \mathrm{Mean}_{i\in \mathrm{Positions}}\ (\mathit{Distance}(\mathrm{BB}i,\ \mathrm{DD}i))$). For each run, we computed the similarity between TL and RL tokens. Across the runs, the distances between TL pairs were significantly smaller than the distances between RL pairs, $t\,(9) = 431.18$, $p < .001$). Thus, under the analogy specified, the Basic Model had a TL effect geometry like the Rueckl et al. model.

3. Interpretation of basic model

3.1. *Depiction of the hidden space geometry*

Next, we sought a visualisation of the geometry of the Basic Model's hidden unit representations. We used principal component analysis (PCA) to analyse the deployment of the 36 input patterns in the 10-dimensional hidden unit space. The first two principal components accounted for more than 98% of the variance in each of the 10 runs. This indicates that virtually all the important structure in the solution lay in the space defined by these two components. Figure 2 shows the hidden positions of all words in this space for one of the

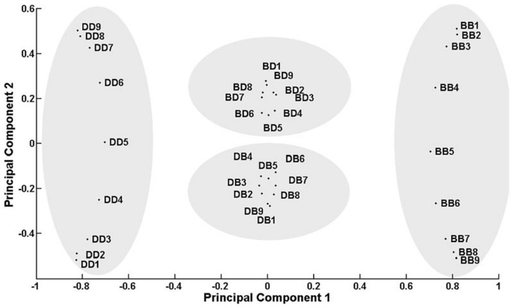

Figure 2. Hidden unit activations of the 36 input patterns projected onto their first two components. The number attached on the right side of the word indicates its input position. For example, BB3 indicates the word BB was presented on the third position which activates the input units, B3 and B4.

runs. Every run showed the same pattern (allowing 180° rotation of the axes). In particular, the following properties held in each of the 10 runs:

(1) Component 1 distinguishes the pure cases (BB vs. DD) and Component 2 distinguishes the mixed cases (BD vs. DB). Each shaded area covers one of the four possible phonological combinations. If we adopt a different basis within the space of Components 1 and 2, referring to the line through $(-1, -1)$ and $(1, 1)$ as the X' axis and the line through $(-1, 1)$ and $(1, -1)$ as the Y' axis, then X' codes B vs. D in the first output slot, while Y' codes B vs. D in the second output slot. (Note that this encoding parallels the kind of featural encoding that is often used in symbolic models.)

(2) Within each quadrant, the set of tokens shows high physical correspondence: that is, the distances between tokens in hidden unit space are roughly proportional to their distances in physical space.[2] Note, for example, that even pairs like BB1 and BB3 that share no input features but are physically nearby tend to be relatively near each other in hidden space; on the other hand physical correspondence is not maximal in any of the groups: the pure classes show modest curvature relative to physical space and the mixed classes lie on ring-like structures. Below, we examine an explicit measure of physical correspondence in order to assess the conditions under which encodings that mirror the geometry of physical space, at least to some degree, arise.

(3) The deployment of tokens within the BB class is opposite the deployment within the DD class (e.g. BB1 is kitty-corner to DD1). In all classes, the peripheral tokens in physical space (e.g. BD1, BB9) are more peripheral in hidden space than the central tokens in physical space (e.g. BD5, BB5). Moreover, the mixed classes are aligned with the pure classes so that each pure class token (e.g. BB9) is closer to the mixed class token that shares its more peripheral letter (DB9) than to the mixed class token that shares its less peripheral letter (BD9).

(4) The 36 patterns have rectangular symmetry (symmetry about a horizontal axis and about a vertical axis). Component 1 (horizontal) is larger than Component 2 (vertical). Moreover, along Component 2, the most peripheral pure cases are more peripheral than the most peripheral mixed cases.

We now consider the reasons for these encoding properties, asking, first, why this geometry exhibits the TL-RL asymmetry, and then why the geometry exhibits a high degree of physical correspondence within classes.

3.2. *Cause of the TL-RL asymmetry*

One source of structure in the hidden unit representations of a feedforward network is the similarity structure of the target patterns. When our network was initialised with small random weights, the hidden unit encodings of all the inputs were distributed around a single point in a normal distribution with small variance. Backpropagation encourages the hidden unit representation of each training exemplar to move toward a common location for exemplars of its own class (i.e. with the same target) and to move apart from exemplars of other classes. The velocities of these movements apart during the training process are partially determined by the similarity structure of the target patterns. The four possible target patterns, along with their inter-pattern distances are shown in Table 1.

Table 1. The inter-target distances between patterns.

Target activations	Pattern name	BB	BD	DB	DD
1 0 1 0	BB	0	$\sqrt{2}$	$\sqrt{2}$	2
1 0 0 1	BD	$\sqrt{2}$	0	2	$\sqrt{2}$
0 1 1 0	DB	$\sqrt{2}$	2	0	$\sqrt{2}$
0 1 0 1	DD	2	$\sqrt{2}$	$\sqrt{2}$	0

Table 2. The distances between the centres of masses of the inputs of the four classes.

Pattern name	BB	BD	DB	DD
BB	0	0.4714	0.4714	0.9162
BD		0	0.2222	0.4714
DB			0	0.4714
DD				0

Since the tokens of BB, BD, DB, and DD, respectively, have identical targets, the hidden unit encodings of the inputs in any one of these classes tend to move in the same direction. With respect to the separation of the members of different classes, the greatest inter-target distances are between BB and DD on the one hand, and BD and DB on the other, so these groups tend to move farthest apart from one another.[3] However, there is no asymmetry in the output structure that could account for the mixed case (BD/DB) vs. pure case (BB/DD) asymmetry that we are interpreting as analogous to TL/RL asymmetries. We turn, therefore, to the structure of the inputs, the only other systematic source of asymmetry in the training set.

The structure of the input patterns also influences the deployment of hidden encodings. When input patterns share features, they tend to produce similar outputs. Therefore, overlapping input patterns which map to different outputs result in contradictory (opposite sign) changes in the hidden locations. On average, then, distinct classes with more shared input structure will separate more slowly than distinct classes with less shared input structure. The amount of shared input structure can be inferred from the distances between the average vectors of each class (closer average vectors imply more shared structure). Table 2 shows the distances in input space between the average vectors of the four classes.

Table 2 makes one source of the TL/RL asymmetry clear: the distance between the average vectors of the two mixed classes (0.22) is lower than that between the average vectors of the two pure classes (0.92). This difference arises because the common letters of the mixed classes coincide in words that are displaced by one letter-position from each other (e.g. BD1 and DB2 share a “D” in slot 2); on the other hand, none of the instances of BB have any overlap with the instances of DD. On average, this makes the BD and DB classes separate more slowly than the BB and DD classes.

In sum, the analysis based on average input vectors reveals a reason for the asymmetry between the BB/DD separation on the one hand, and the BD/DB separation on the other, which is at least partially responsible for the TL/RL asymmetry in the Basic Model. This analysis does not, however, address the within-class physical correspondence that we remarked on above.

3.3. Cause of high physical correspondence within classes

As noted above, the Basic Model seems to exhibit high within-class physical correspondence. This property is interesting because, as noted above, there is no simple indicator of spatial position in the inputs or the outputs to the model. This section of the paper first quantifies the notion of "physical correspondence" and then probes the reason for high within-class physical correspondence. The section will show that the cause of the within-class physical correspondence in the Basic Model is a kind of "domino-effect" in the learning process which is triggered by what seems like a representationally unimportant asymmetry in the structure of the input, allowing us to pinpoint the reason that the Basic Model encodings both resemble physical spatial structure and diverge from it. In General Discussion, we explain how this analysis allows us to conclude that the Basic Model is an example of a perceptual system which is not decomposable into a symbolic level of description and an implementational level of description (Fodor & Pylyshyn, 1988), hence providing an explicit demonstration of a connectionist encoding that is non-classical. The section will also show that the Rueckl et al. model shows relatively high within-class physical correspondence, suggesting that the analysis of causes provided here may apply in a more realistic model as well.

3.3.1. Quantification of "Physical correspondence"

The spatial structure "in the world" that interests us here is the physical deployment of written words on a page. We are interested not in absolute spatial position but in position relative to an observer's point of focus. For simplicity, the current model assumes that input spatial positions differ only in one spatial coordinate, which is analogous to the "horizontal coordinate" in normally oriented perception of written stimuli in a language like English, which is written from left to right with spaces between the words.[4] It is clear by inspection that the model does not form a hidden code that is a linear scaling of the physical positions of tokens. But within classes of tokens mapping to the same output, physically nearby tokens tend to lie approximately along straight lines in hidden unit space. We can estimate the degree of this approximation by considering the average angle between the line segments connecting the encodings of physically adjacent codings in hidden unit space:

$$\mathrm{PhysCorr} = \frac{1}{(n-2)} \sum_{i=1}^{n-2} \mathrm{angle}(\overline{t_{c,i}t_{c,i+1}}, \overline{t_{c,i+1}t_{c,i+2}}), \tag{3}$$

where $t_{c,i}$ is the hidden unit encoding of the i'th token of word c (e.g. $t_{\mathrm{BB},1}$ is the hidden unit position of BB1), $\overline{xy}$ is the line segment connecting points x and y, $angle(\overline{xy},\overline{yz})$ is the angle sweeping segment $\overline{yz}$ into segment $\overline{xy}$, and n is the number of tokens in the class. A set of hidden unit locations exhibiting maximal physical correspondence with the linear sequence of input locations has PhysCorr = 180°. By Monte Carlo simulation, we discovered that a collection of points distributed randomly (normally) in a space of any dimension visited in random sequence has PhysCorr close to 60°. Thus, to assess degree of physical correspondence, we should consider observed values in relation to these two reference points.

Row 1 of Table 3 shows physical correspondence values for each of the four classes in the Basic Model. For comparison, Row 2 shows physical correspondence for the Orthogonal

Table 3. Mean (standard deviation) of physical correspondence of different models.

	BB	BD	DB	DD	Mean
Basic model	141.82 (1.73)	105.64 (8.65)	110.10 (6.65)	140.98 (2.39)	124.64 (3.74)
Orthogonal inputs model	71.09 (25.04)	57.00 (32.34)	69.08 (28.18)	55.55 (23.87)	63.18 (19.52)
No overlap model	73.28 (14.99)	72.91 (8.90)	68.16 (15.60)	66.76 (14.31)	70.27 (9.45)
Ring Model	99.01 (17.46)	75.65 (21.92)	83.93 (13.51)	110.02 (23.10)	92.15 (9.87)
Rueckl et al. (2008)					71.13 (4.86)

Notes: The PhysCorr values were measured in the full hidden unit space (not, e.g., the space of the dominant principal components) in each case. See the text for descriptions of the various models. Values are in degrees. Standard deviations are shown in parentheses. The mean value for the Rueckl et al. model was computed by computing a mean for each word (across all positions) and then computing the means across these means. The standard deviation shown in the Mean column gives the variation across means/runs (ignoring within-class variability).

Inputs Model mentioned in Footnote 3 above. The Orthogonal Inputs model has no structure in the input space (the inputs of all 36 tokens are orthogonal, equal-length vectors), so there is no information about physical location in the input. Thus, the PhysCorr values are close to the mean random value. The No Overlap Model (Row 3) is a related case: in this model, the input coding is the same as in the Basic Model, but only every other token is used (there are 40 input units and 9 tokens per class). In this case, there is no overlap between adjacent words (the only overlap in the input encodings occurs when words in the same position have the same first letter or the same second letter). Again, not surprisingly, since there is no information in the input or the output specifying information about the proximity relationships of words, the PhysCorr values are near the mean random value. We also computed the mean PhysCorr value from Rueckl et al. model (bottom row). The value is not far from the mean random value and similar to the Orthogonal Inputs Model and the No Overlap Model, although the standard deviation is smaller than in both these cases.

The low value in the Rueckl et al. model seems, at first glance, to indicate that physical correspondence is not a feature of the positional encodings in that model. However, further analysis suggests a different view. The pattern of input similarity in that model is much more complicated than in our model, because the 2998 English words used to train that model have letters in common in many different positions (not just adjacent positions). When we measured PhysCorr in the Rueckl et al. model for just those words that have adjacent double letters (e.g. "add", "eel"), the value (mean $= 79.65$, SD $= 4.43$) was significantly higher than length and frequency matched, non-double-letter controls ($t(358) = 36.15$, $p < .0001$). Moreover, although words with double letters separated by one letter (e.g. "stitch", "pipe") showed averages close to the minimum (mean $= 63.23$, SD $= 2.82$), when we computed angles using every other position (e.g. stitch1–stitch3–stitch5, stitch2–stitch4–stitch6, etc.), the mean PhysCorr was significantly higher than for frequency matched controls (mean $= 76.44$, SD $= 3.90$; $t(61) = 33.20$, $p < .0001$).

These results suggest that the Rueckl et al. model exhibits physical correspondence of its within word encodings under conditions parallel to those where we observed it in the Basic Model, thus providing a motivation for looking more closely at the cause of the physical correspondence.

3.3.2. Causes of physical correspondence

What is the source of the within-class physical correspondence in the Basic Model? To answer this, it is helpful to understand how the learning process drives the model into the hidden space configuration in Figure 2.

First, as noted above, the velocity of separation of each target class is related to the average distance between members of the class. Since the pure classes (BB and DD) are more distant, on average, from each other, than the mixed classes (BD and DB), we expect separation along the first principal component (corresponding to the B vs. D contrast) to happen earlier than separation along the second principal component. Indeed, every Basic Model simulation showed this pattern.

Now, consider an idealised weight set that approximates a learning stage when the pure cases have separated but the mixed cases have not. For clarity of comprehension, it is convenient to think of an "Idealized Network" with only two hidden units, the hidden units corresponding, respectively, to the first and second principal components in Figure 2. Thus, we assume that the weights from all D units to the first component are negative and from all B units are positive (and all these are equal in magnitude). At this point, we also assume that the training has produced an appropriate mapping from the first hidden unit to the outputs (turns on DD when the first hidden unit is negative and BB when the first hidden unit is positive). We also assume that all the weights from input to the second hidden unit are zero (no separation along the second component yet) but that the initial weight randomisation has produced an asymmetry in the output units that favours turning BD on when the second hidden unit is negative and DB on when the second hidden unit is positive. For concreteness, and without loss of generality, we can assume these weights have the values shown in Tables 4 and 5.

When the Idealised Network is trained starting from this configuration, some of the weights from the input to the second hidden unit will diverge from zero. In particular, the average signal to the weights associated with peripheral positions is biased. For example, the weight H2-B1 is adjusted only when the patterns BB1 and BD1 are presented. We may assume that the activation function gain is high enough that BB1 produces no significant error at this point. Therefore, the only force on H2–B1 is due to BD1 and this is a negative force (due to the polarity of the hidden → output mapping). There is an equal negative

Table 4. Weights from input to hidden units (columns index inputs, rows index hiddens) in the Idealised Network when the first component has separated but the second has not.

	1 B1	2 D1	3 B2	4 D2	5 B3	6 D3	...	19 B10	20 D10
H1	1	−1	1	−1	1	−1		1	−1
H2	0	0	0	0	0	0		0	0

Note: H1 and H2 are first and second hidden units.

Table 5. Weights from hidden to output units (columns index hiddens, rows index outputs) in the Idealised Network when the first component has separated but the second has not.

	H1	H2
1 B1	1	−1
2 D1	−1	1
3 B2	1	1
4 D2	−1	−1

Note: H1 and H2 are first and second hidden units.

force on H2–D10 (from the pattern BD9), and corresponding (equal) positive forces on H2–D1 and H2–B10. All of the other influences on the weights from input B2 add to zero because of the symmetric behaviour of the non-peripheral units (e.g. H2–B2 receives a positive signal from DB1 and an equal, negative signal from BD2). Thus, on the first epoch of training, only the peripheral weights are adjusted. On the second epoch of training, however, the adjustment made on the first epoch changes the balance of forces on the adjacent-to-peripheral units (B2, D2, B9, D9). Essentially, the progress toward separation of the peripheral units results in reduction in the error on peripheral (mixed) patterns, so the adjacent-to-peripheral units start to undergo error changes as though they were peripheral units, though more weakly than the actual peripheral units (Appendix 1). We note that this chain of successive causes and effects has the form of a "domino process": one event triggers the next, which triggers the next, etc. Since the changes are all symmetric with respect to first vs. second position, and also symmetric with respect to the B/D contrast, the following properties hold:

(1) $w_{\mathrm{h2-B}i} = -w_{\mathrm{h2-D}i}$ for $i = 1 \ldots 9$
(2) $w_{\mathrm{h2-D}i}$ is monotone decreasing and $w_{\mathrm{h2-B}i}$ is monotone increasing in *i*.

Moreover, when the number of epochs reaches half the number of positions, the monotonicity becomes strict, and remains so throughout training. Conveniently, with strict monotonicity, Properties 1 and 2 above suffice to separate the mixed cases. Figure 3 shows the development over training of the input to second hidden unit weight pattern in the Idealised Network. Alluding to Property (1), we refer to this pattern as the "Additive Inverse Solution". To see why the Additive Inverse Solution suffices to separate the mixed cases, let $f(i) = w_{\mathrm{h2-D}i}$ and $g(i) = w_{\mathrm{h2-B}i}$. Then, for $\Delta > 0$,

$$f(i) = -g(i),$$
$$f(i) + g(i) = 0,$$
$$f(i) + g(i + \Delta) > 0$$

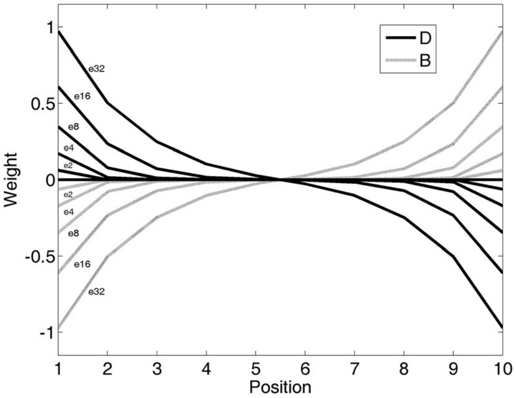

Figure 3. Weight development in the Idealised Network. Each curve shows weight $w_{\mathrm{h2-(B/D)}i}$ for *i* ranging across positions (the depicted run employed only 7 positions B1D1 ... B7D7). The labels *ej* specify the epoch numbers.

because $g(i)$ is strictly increasing and

$$f(i+\Delta)+g(i)<0$$

because f is strictly decreasing. Since the patterns DB are all of the form $f(i)+g(i+\Delta)$ (e.g. DB1 $=$ D1 $+$ B2, $\Delta=1$) and the patterns BD are all of the form $f(i+\Delta)+g(i)$, the summed functions separate the patterns.

This analysis, formalised in Appendix 1, predicts the U-shaped physical correspondence in the second principal component exhibited by the mixed cases. It also predicts the linear physical correspondence in the second component exhibited by the pure cases. As predicted by this analysis, when the Idealised Network was trained to zero error, it converged on the hidden space structure shown in Figure 2 (and described in points i through iv) and exhibited the Additive Inverse Solution shown in Figure 3. If the Basic Model is operating on the same principles as the Idealised Network, then we would expect the second component of the hidden unit displacements to approximate Figure 3 pattern. Indeed, each run of the Basic Model showed such a pattern (Figure 4).

What does the analysis reveal about the cause of the pattern of encoding? Two points are important: First, the physical correspondence stems only indirectly from the physical configuration of the input – the domino process capitalises on the shared sensory nodes of words in adjacent positions, leading to a progressive expansion of the use of the second encoding component (H2 in the Idealised Network). This gives rise to an encoding that exhibits partial, but not total physical correspondence – adjacent members are encoded near each other, but there is curvature in the mixed cases such that XYk and XY$(9-k+1)$ have the same second component value.[5] Second, the discovery of the Additive Inverse solution depends on the existence of a representationally-irrelevant asymmetry in the input (the contrast between peripheral and central positions). We take up these points in the General Discussion, considering their implications for distinguishing connectionist and classical treatments of the nature of perception.

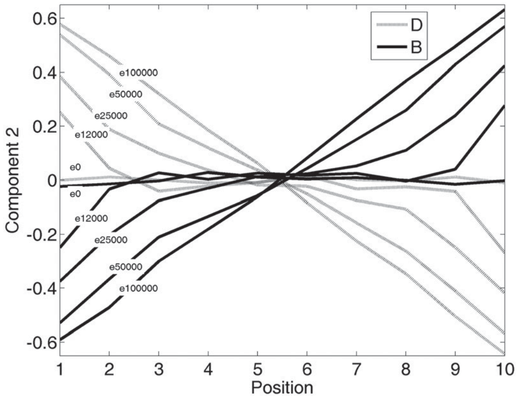

Figure 4. Weight development in the basic model. Each curve shows a letter across positions projects onto the second component. The labels *ej* specify the epoch numbers.

3.4. Testing the role of asymmetry: the Ring Model

An interesting further prediction of this analysis is that, if there is no distributional asymmetry between the constituents of the peripheral and non-peripheral cases, then the model is very unlikely to develop the additive inverse solution. This is because the additive inverse solution depends on a symmetry of the second component weights ($w_{\mathrm{h2-B}i} = -\mathrm{w}_{\mathrm{h2-D}i}$) Without the structuring provided by the physical asymmetry of the input distributions, the network has to rely on random processes (the initial weight setting and the random sequencing of the patterns during training) to discover the relevant weight symmetry. Since 20 weights are involved, the chances of producing an effective approximation of this configuration randomly are exceedingly small.

To test the importance of the asymmetry, we designed the Ring Model. This model has only 18 input units. All of the input patterns are the same as in the basic model, except that the patterns labelled XY9 have inputs of the form [Y1, X9] (for example, BD9 has D1 and B9 on, all other units off). Thus, the inputs to this model lie on a ring and they have no peripheral-central asymmetry. As predicted, this model develops hidden unit separation along a component that distinguishes B from D, but it fails to develop a significant second component. Not surprisingly, the model fails to completely learn the map: it predicts the pure cases well, but fails to distinguish the BD from the DB mixed cases. Moreover, as shown in Table 3, the model fails to develop within-class physical correspondence at levels comparable to those of the Basic Model.

4. General discussion

4.1. Summary

We have described a feedforward neural network that learns to classify structured objects position independently. The objects are simple analogs of word stimuli. Two results stand out: the model exhibits an analog of the empirically attested TL–RL asymmetry and it exhibits within-class physical correspondence – that is, within the classes of stimuli defined by the task, the model employs a context-sensitive encoding that bears a resemblance to the spatial organisation of the stimuli themselves. Our analyses helped show how these features arose through the interaction of the learning mechanism with the data. Now we discuss implications of these findings for theories of written word perception and of mental encoding more generally.

4.2. TL–RL asymmetry and theories of written word perception

Regarding TL–RL asymmetry, the analysis shows that the greater similarity between individual TL pairs than RL pairs arises because of the overall similarity of the classes to which these individual pairs belong. Specifically, the class of BD words has more letter-instantiations in common with the class of DB words than the class of BB words has in common with the class of DD words. This property stems from the functional structure of the task: the fact that the same physical object (e.g. D2) plays one role (onset) in one class (DB) and a different role (coda) in another class (BD). Thus, the predictions of the current framework differ from those of frameworks that claim that the TL-RL asymmetry is the result of uncertainty in the perception of position (e.g. Davis & Bowers, 2006; Gomez et al., 2008; Grainger & Van Heuven, 2003; Whitney, 2001; see Grossberg, 1978) and is independent of the functional structure of the task. The functional hypothesis, but not the uncertainty hypothesis, predicts that TL-RL effects should be reduced if the classes share less structure. The functional hypothesis is thus consistent with the results of Lee and Taft (2009) who found that TL-RL effects are much weaker for readers of Korean Hangul, where onsets and codas are positioned systematically differently within each written character. The functional hypothesis also suggests that letter transposition effects should be sensitive to the structural roles that letters play – for example, confusability should be different if elements are exchanged across similar roles (e.g. exchanging two consonants) as opposed to across contrasting roles (e.g. exchanging a vowel and a consonant). The positional uncertainty view expects physical proximity to increase similarity regardless of structural role.

4.3. The status of symbolic theories

We asked at the beginning, whether the recent theories, which emphasise contextually specific mental encoding, require a reformulation of symbolic theories of mental representation, or if there is a trivial fix, via, for example, the notion of peripheral resonance. To make our arguments explicit, we focus on a version of the symbolic theory whose core is compositionality. Based on the findings reported in the body of the paper, we argue that, for the model and environment considered here, while symbolic insights are partially relevant, the contextual dependence observed is of a kind that does not fit within this symbolic framework. These considerations do not argue for or against the claim that humans employ symbolic computation, but they help make precise an alternative to the symbolic account.

We focus, for clarity, on one representative of the symbolic view: model theoretic syntax/semantics (Dowty, Wall, & Peters, 1981; Heim & Kratzer, 1998). This view assumes that there is a syntax that specifies ways in which complex symbols are built out of other complex symbols or primitive symbols by concatenation. Both kinds of symbols correspond to sets of entities in the world. For each rule of concatenation, there is a rule of semantic interpretation which says how the set in the world corresponding to the complex symbol the rule produces is derived from the sets in the world corresponding to the constituent symbols. This rule provides the semantics for all symbols that can legally be constituents in the rule.

When we say that this framework is the "essence" of the symbolic theory, we are alluding to claims that the model theoretic framework is all that cognitive science needs to care about. For example, (Fodor & Pylyshyn, 1988) note that an appealing property of the framework is that the syntactic system affords physical implementation (in logic gates), but the details of the physical implementation are independent of the structural principles just outlined; in other words the framework is multiply realisable. This view amounts to a profound form of modularity – it says that a line can be drawn between the "implementation" level and the "compositional" (or "cognitive" or "representational" or "symbolic") level[6] such that all the insights that cognitive science needs to discover are insights about the organisation of the compositional level. The implementation level can be ignored.

An interesting property of the Basic Model is that part of it (the hidden-output layer weights) implement a compositional system. In particular, one dimension (X′) of the hidden space visitation set is a variable whose (two) values are primitive symbols. These primitive symbols in the "syntax" of the network's mental system correspond, respectively, to "words with B in the first slot" and "words with a D in the first slot". Such words are objects in the world that the network inhabits. An orthogonal dimension (Y′) similarly "represents" the second slot.

What, then, of the input-hidden weights? On Fodor and Pylyshyn's story, these weights should either be another part of the compositional system which also exhibits canonical intentionality, or they should be part of the "implementation" and thus not relevant to explaining cognition. We will argue that the input-hidden map does not have compositional syntax and semantics and yet a theory of the network's mental system can hardly afford to ignore it.

There are three points:

(1) The input-hidden map has similarities to a compositional system, but it does not appear to be one. For example, let us assume that the model is implemented as a real device for providing what amount to "phonological transcriptions" of written words shown in various physical positions. The within-class physical correspondence in the hidden space, which stems from the form of the input-hidden map, looks like a kind of "representation" of the positions of words in the world. It could even be used to make some correct inferences – for example, that a small shift in gaze direction will turn BB6 into BB5. However, the extension of this inference to the mixed peripheral cases yields some erroneous conclusions (e.g. BD1 is adjacent to BD9 in hidden space, but not in the world). More generally, the mixed class physical correspondence is different from the pure class physical correspondence. These observations cast doubt on the hypothesis that the within class similarities are representations. Moreover, the analysis has shown that these similarities, such as they are, exist because of the role they play in mapping orthographic forms to phonological encodings, which, as we have noted, has a representational structure in the form of slot codes. If the input-hidden map happened to produce a perfect (unwarped) physical correspondence in support of the hidden-output compositional map, then we would be justified in claiming that "representation" is an accurate description of both levels. This is the way researchers have often thought about models like McClelland and Rumelhart's (1981) Interactive Activation model of word and letter perception: the model "represents" the world simultaneously at the feature level, the letter level, and the word level. However, in the present case, given that the warped similarities arose in support of the hidden-output representational system, it is not clear that we should consider the input-hidden map representational.

(2) As shown by the analysis, TL-RL effects (as construed in this network) depend on both the asymmetry of the principal components (first larger than second) and the inverse relationship between the two pure class similarities. Both of these features stem from the structure of the input-hidden map. If we take the analogy with real word perception at face value, then the encoding differences are responsible for the fact that people are more likely to perceive a TL stimulus as the word it was derived from than they are to perceive an RL stimulus as such. Although we, as researchers, and the perceivers themselves (on more careful inspection) may interpret such behaviours as "errors", they are nevertheless an aspect of our perceptual structure that the theory of cognition may do well to attend to. The case would be different if the errors had a random distribution, with TL and RL stimuli equally likely to be erroneously mistaken for words – then

the claim that the classical model fully describes the structure of perception would be reasonable – the error-tendencies could be treated as a uniform fuzzing of the structural picture, not requiring any nuance (e.g. dimensional asymmetry/curvature) in our account of the correct encodings.

(3) The analysis showed that the network's discovery of the representational scheme of the hidden-output weights depended on a physical asymmetry in the input structures. In particular, multiple Ring Model simulations, in which the physical asymmetry was absent, failed to discover a solution to the map. One might wonder if the Ring Model's failure stems from an architectural inability of the model to solve the Ring Task. It does not: Appendix 2 gives an example of a weight set that solves the Ring task. The analysis of the Basic Model showed that, to solve the Ring Task with a system derived from that of the Basic Model, the system would have to break the initial symmetry of the input-hidden weights in a particular way that would be highly unlikely to happen by chance. This suggests that the physical asymmetry of the inputs, which spurs the discovery of the delicate symmetry breaking in the Basic Model is an important element of the model's facility for knowledge discovery. This contrasts with Fodor and Pylyshyn's arguments: if the computational/compositional level of analysis is independent of the peculiarities of implementation, the input asymmetries should play no role. If we were to adopt their position, we would run the risk of overlooking this critical detail of how the Basic Model solves the perceptual task and its implications for the form of the hidden-layer encoding.

We remark, by way of conclusion, that this detail may open a helpful avenue of future investigation. It is challenging to get neural feedback systems to learn complex, symbolically structured tasks (see Bengio, Simard, & Frasconi, 1994; Tabor, 2003; Tabor, Cho, & Szkudlarek, 2013). The feedback interaction between the network and the environment must drive the input-hidden weights to undergo complex, coordinated role-differentiation.[7] The randomisation of the initial weights seeds this role differentiation. Whenever a random process assists in structure discovery, we should ask to what extent the random process is doing the work of the discovery. If, in fact, the random weight seed were doing all the work of endowing the input-hidden map with a suitable topology, and the training simply served to configure the hidden-output weights to take advantage of this topology, then there would be little to be gained by studying the geometry of the network encodings. But, in the present case, as we noted in the Idealised Network investigation, the weight randomisation only has to make a very simple choice – it has to decide whether BD should be positive and DB negative, or vice versa, a very easy choice for a random process. Once that choice is made, the interaction of the network with the training environment drives the domino process to reach an effective encoding.

This observation suggests asking for what learning tasks there are domino pathways to the construction of complex symmetries. The answer to this question can be positive even in cases where the network fails to learn: we seeded a version of the Ring Network with small values proportional to those given in Appendix 2 for just the weights B1-1, B1-2, B1-3, D1-1, D1-2, D2-4, B10-1, B10-2, B10-4, D10-1, D10-2, D10-3 (a total of 12 weights), and the network reliably created appropriate differentiation in the other 72 weights and biases governing the hidden units, thereby succeeding with the Ring Task. Even though guessing this 12-weight configuration by randomising the initial weights is very unlikely to be successful, the fact that it exists indicates that a set of domino-pathways to solution (at least 10 of them, by symmetry) lies close to the unbiased initial state (all weights zero). These pathways exist because of a symmetry in the problem: changing the position of a word does not change its pronunciation; a useful research avenue then, may be to establish techniques of discovering such domino pathways from problem symmetries.

Summarising, the considerations mentioned above provide some impetus against the strong modular perspective and in favour of what might be called the "full physical picture" perspective: cognitive science needs to take seriously the physical nature of organisms and their environments; although the classical symbolic view is committed to physical instantiation, it overlooks a possible role that variation in physical instances not directly relevant to those abstractions plays in the discovery of those abstractions and in determining the structure of mental encodings. The models discussed here also offer examples of how something that may look, at first glance, like a brain simulacrum of environmental structure may not be precisely that (recall the curvature across position encodings), and may arise in an indirect way from a regularity in the environment. Finally, building on the suggestion that symbols are not atomic but have a complex internal structure with complex gestation, the results suggest pursuing an understanding of complex learning by studying "domino pathways" of cause and effect in the learning process.

Notes

1. Mental simulation of the world should not be confused with *perceptual simulation* (Barsalou, 1999), which refers to conceptual representations that are simulations of perceptual representations.
2. We use the term "similar" to mean "same in form". Whereas in geometry, similarity is a binary property (two shapes are either similar or not), we refer here to degrees of similarity. Below, we provide a quantitative definition in order to measure degrees of physical correspondence in particular models.
3. The lowest dimensional point-set that exhibits the inter-target distances in Table 1 is a square with BB and DD at diagonally opposite vertices and BD and DB also diagonally opposite. In fact, if we replace the actual inputs of the model with the indexical bit vectors in R – that is, a unique unit with activation one and the rest zero for each input – an encoding which lacks asymmetries in the input space, then train that model (the "Orthogonal Inputs Model"), the hidden unit pattern closely approximates a square with this structure. This square is roughly similar to the hidden pattern of the Basic Model (as shown in Figure 2), but it lacks two properties of central interest here: a TL/RL asymmetry and physical correspondence of the within-class encodings.
4. The current model ignores the fact that people can read words displaced above and below as well as to the left and the right of their focus, that they can perceive words in non-horizontal orientations, and, as noted above, that they can read them in continuously varying positions, at least within a small visual angle surrounding the focal point.
5. The reader may have noted that there is also curvature in the first component. This also distorts the encoding away from perfect physical correspondence. This curvature may occur in part because the separation of DB and BD on the second component causes the pure cases to array themselves oppositely on this component, and this, in turn creates opportunities for elements to reduce error by adjusting on the first component, but we do not, at present know why the first component curvature takes precisely the form it does.
6. In the terms of Marr (1982), we are concerned here with the separation between the "implementation" and the "algorithmic" level.
7. This process is an instance of what is called *self-organisation* – cases where continuous feedback interactions among many small, interacting elements give rise to organised structure at the level

of the ensemble (Haken, 1983; Johnson, 2002; Koschmieder, 1993; Kukona, Cho, Magnuson, & Tabor, 2014; Kukona & Tabor, 2011; Tabor & Hutchins, 2004; Tabor, Galantucci, & Richardson, 2004; Zhabotinsky, 1991). Self-organisation is plausibly central in the formation of complex mental structure, but a general formal theory of self-organisation is lacking (see Bak, 1996; Jelinek, 1990 for relevant forays in physics). The phenomenon we refer to here by the term "domino pathway" may be a useful focus point for a general theory of self-organisation.

Acknowledgements

Preliminary results of this modelling were presented at the 4th University of Connecticut Dynamics and Cognition Workshop (August, 2009) and the 4th Computational Cognitive Neuroscience Conference (November, 2009).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the NSF under Grant [INSPIRE 1246920] to the University of Connecticut via the University of Illinois; National Institute of Child Health and Human Development under Grant [HD40353] to Haskins Laboratories; and National Institutes of Health under Grant [P01 HD 001994] to Haskins Laboratories. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Acha, J., & Perea, M. (2008). The effect of neighborhood frequency in reading: Evidence with transposed-letter neighbors. *Cognition*, *108*, 290–300. doi:10.1016/j.cognition.2008.02.006

Bak, P. (1996). *How nature works: the science of self-organized criticality*. New York, NY: Copernicus Press.

Barsalou, L. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*, 1177–1187. doi:10.1098/rstb.2003.1319

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*, 84–91. doi:10.1016/S1364-6613(02)00029-3

Beauchamp, M. S. (2005). See me, hear me, touch me: Multisensory integration in lateral occipital-temporal cortex. *Current Opinion in Neurobiology*, *15*, 145–153. doi:10.1016/j.conb.2005.03.011

Beauchamp, M. S., Lee, K. E., Haxby, J. V., & Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron*, *34*, 149–159. doi:10.1016/S0896-6273(02)00642-6

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*, 157–166. doi:10.1109/72.279181

Bryson, A. E., & Ho, Y. C. (1975). *Applied optimal control: Optimization, estimation, and control*. New York: Hemisphere.

Chambers, S. M. (1979). Letter and order information in lexical access. *Journal of Verbal Learning and Verbal Behavior*, *18*, 225–241. doi:10.1016/S0022-5371(79)90136-1

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Davis, C. J., & Bowers, J. S. (2004). What do letter migration errors reveal about letter position coding in visual word recognition? *Journal of Experimental Psychology*, *30*, 923–941. doi:10.1037/0096-1523.30.5.923

Davis, C. J., & Bowers, J. S. (2006). Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. *Journal of Experimental Psychology*, *32*, 535–557. doi:10.1037/0096-1523.32.3.535

Dowty, D. R., Wall, R. E., & Peters, S. (1981). *Introduction to Montague semantics*. Dordrecht: D. Reidel.

Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*, 183–204. doi:10.1016/0010-0277(90)90014-B

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71. doi:10.1016/0010-0277(88)90031-5

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *39*, 211–251. doi:10.1080/14640748708401785

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin and Review*, *9*, 558–565. doi:10.3758/BF03196313

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review*, *115*, 577–600. doi:10.1037/a0012667

Grainger, J., & Van Heuven, W. (2003). Modeling letter position coding in printed word perception. In P. Bonin (Ed.), *The mental lexicon* (pp. 1–23). New York: Nova Science.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen & F. Snell (Eds.), *Progress in theoretical biology* (vol. 5, pp. 233–374). New York, NY: Academic Press.

Haken, H. (1983). *Synergetics, an introduction: Nonequilibrium phase transitions and self-organization in physics, chemistry, and biology*. New York: Springer-Verla.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720. doi:10.1037/0033-295X.111.3.662

Heim, I., & Kratzer, A. (1998). *Semantics in generative grammar* (vol. 13). Oxford: Blackwell.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, *40*, 185–234. doi:10.1016/0004-3702(89)90049-0

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, *1*, 295–307. doi:10.1016/0893-6080(88)90003-2

Jelinek, F. (1990). Self-organized language modeling for speech recognition. In A. Waibel & K. F. Lee (Eds.), *Readings in speech recognition* (pp. 450–506). San Mateo, CA: Morgan-Kaufmann.

Johnson, J. S., Spencer, J. P., & Schöner, G. (2008). Moving to higher ground: The dynamic field theory and the dynamics of visual cognition. *New Ideas in Psychology*, *26*, 227–251. doi:10.1016/j.newideapsych.2007.07.007

Johnson, S. (2002). *Emergence: The connected lives of ants, brains, cities, and software*. New York: Simon and Schuster.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Koschmieder, E. L. (1993). *Bénard cells and Taylor vortices*. Cambridge: Cambridge University Press.

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 326–347.

Kukona, A., & Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. *Cognitive science*, *35*(6), 1009–1051.

Lee, C. H., & Taft, M. (2009). Are onsets and codas important in processing letter position? A comparison of TL effects in English and Korean. *Journal of Memory and Language*, *60*, 530–542. doi:10.1016/j.jml.2009.01.002

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, *102*, 59–70. doi:10.1016/j.jphysparis.2008.03.004

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: W.H. Freeman and Company.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45. doi:10.1146/annurev.psych.57.102904.190143

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology*, *11*(2), 194–201. doi:10.1016/S0959-4388(00)00196-3

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*, 287–330. doi:10.1037/0033-295X.86.4.287

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception: Part 1. *Psychological Review*, *88*, 375–407. doi:10.1037/0033-295X.88.5.375

McClelland, J. L., Rumelhart, D. E., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II*. Cambridge, MA: MIT Press.

Michaels, C. F., & Carello, C. (1981). *Direct perception*. Englewood Cliffs, NJ: Prentice-Hall.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115. doi:10.1037/0033-295X.103.1.56

Prindle, S. S., Carello, C., & Turvey, M. T. (1980). Animal-environment mutuality and direct perception. *Behavioral and Brain Sciences*, *3*, 395–397. doi:10.1017/S0140525X0000563X

Rueckl, J. G., Fang, S.-Y., Begosh, K. T., Rimzhim, R., & Tobin, S. (2008). *Learned internal representations and letter position information: A connectionist approach*. 49th Annual Meeting of the Psychonomic Society. Chicago, IL, USA.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp. 1–34). Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. doi:10.1038/323533a0

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I*. Cambridge, MA: MIT Press.

Schoonbaert, S., & Grainger, J. (2004). Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language and Cognitive Processes*, *19*, 333–367. doi:10.1080/769813932

Simmons, W. K., Hamann, S. B., Harenski, C. L., Hu, X. P., & Barsalou, L. W. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology-Paris*, *102*, 106–119. doi:10.1016/j.jphysparis.2008.03.014

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–23. doi:10.1017/S0140525X00052432

Smolensky, P. (1991). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. In T. Horgan & J. Tienson (Eds.), *Connectionism and the philosophy of mind* (pp. 281–308). Dordrecht: Springer.

Tabor, W. (2003). Learning exponential state-growth languages by hill climbing. *IEEE Transactions on Neural Networks/a Publication of the IEEE Neural Networks Council*, *14*, 444–446. doi:10.1109/TNN.2003.809421

Tabor, W., Cho, P. W., & Szkudlarek, E. (2013). Fractal analysis illuminates the form of connectionist structural gradualness. *Topics in Cognitive Science*, *5*, 634–667. doi:10.1111/tops.12036

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*(4), 355–370.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 431–450.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge: MIT Press.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin and Review*, *8*, 221. doi:10.3758/BF03196158

Wu, L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, *132*, 173–189. doi:10.1016/j.actpsy.2009.02.002

Zhabotinsky, A. M. (1991). A history of chemical oscillations and waves. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *1*(4), 379–386.

Appendix 1

Consider the simplified network described in Section 3.3.1. Suppose that, at some point in training, the network's weights satisfy the following conditions:

$$\begin{aligned} &w_{\mathrm{OB1-H2}} = w_{\mathrm{OD2-H2}} = w_{\mathrm{oh}} \\ &w_{\mathrm{OD1-H2}} = w_{\mathrm{OB2-H2}} = -w_{\mathrm{oh}} \\ &w_{\mathrm{oh}} \gg 0 \end{aligned} \tag{A1}$$

$$\begin{aligned} &w_{\mathrm{H2-B1}} > w_{\mathrm{H2-B2}} > \cdots > w_{\mathrm{H2-B}k} \\ &w_{\mathrm{H2-D1}} < w_{\mathrm{H2-D2}} < \cdots < w_{\mathrm{H2-D}k} \\ &1 < k < \frac{n}{2} \end{aligned} \tag{A2}$$

$$\begin{aligned} &w_{\mathrm{H2-D}i} = -w_{\mathrm{H2-B}i} \\ &1 \leq i \leq n. \end{aligned} \tag{A3}$$

The weights, $w_{\mathrm{H2–X1}}, \ldots, w_{\mathrm{H2–X}k}$ then satisfy the strict monotonicity condition described in Section 3.3.1. We wish to show that further training will cause the system to completely satisfy the strict complementary montonicity condition, that is, that the above relations will eventually hold for $k = n$ and this situation will hold at all future times). Suppose the error function is structured so that $w_{\mathrm{H2–B}k}$ increases from this point but does not become as large as – $w_{\mathrm{H2–D}(k-1)}$. Then, if $W_{\mathrm{H2–B}k}$ is zero, it will become positive at the next time step. Likewise, by symmetry, $W_{\mathrm{H2–B}(n-k+1)}$ will become equally positive at the next time step, and $w_{\mathrm{H2–D}k}$ and $W_{\mathrm{H2–D}(n-k+1)}$ will become equally negative at the next time step. If these four weight values have already moved away from zero, they will continue to do so, but their absolute value will never reach $w_{\mathrm{H2–B}(k-1).}$ Consequently, after $n/2$ time steps, the relationships (A1)–(A3) will hold for $k = n$ and they will continue to hold throughout training.

If it is also the case that, initially, the following inequalities hold,

$$\begin{aligned} &w_{\mathrm{H2-B1}} < w_{\mathrm{H2-B2}} \\ &w_{\mathrm{H2-D1}} > w_{\mathrm{H2-D2}} \\ &w_{\mathrm{H2-B}n} > w_{\mathrm{H2-B}(n-1)} \\ &w_{\mathrm{H2-D}n} < w_{\mathrm{H2-D}(n-1)} \end{aligned} \tag{A4}$$

then, by mathematical induction, the system will, in finite time, satisfy (A1)–(A3) for $k = n$ and remain in this state through all future time steps.

We assume that all the weights $w_{\mathrm{H2–X}i}$ start out at zero (this approximates the situation in the Basic Model, where these weights start with small random values and these values do not change much during the initial growth of the weights $w_{\mathrm{H1–X}i}$).

A1.1 Base case

To establish the base case (A4), we focus on the first inequality, and note that only the patterns BB1 and BD1 have non-zero activation of B1 and only the patterns BB1, BB2, DB1, and BD2 have non-zero activation of B2. Therefore, we can ignore all other patterns in determining the initial change of $w_{\mathrm{H2–B1}}$ and $w_{\mathrm{H2–B2}}$. The weights from the input to H1 are set so that the network makes essentially zero error on BB1 and BB2. Therefore, BD1 is the only pattern that will significantly influence $w_{\mathrm{H2–B1}}$. The values of the hidden-to-output weights imply (via backpropagation) that training for one time step on BD1, will make $w_{\mathrm{H2–B1}}$ grow positive. Similarly, the only training patterns that induce change in $w_{\mathrm{H2–B2}}$ under the circumstances are DB1 and BD2. This time, the values of the hidden-to-output weights imply equal and opposite changes when DB1 and BD2 are presented (given that $w_{\mathrm{H2–D1}}$ and $w_{\mathrm{H2–D3}}$ are zero). Therefore, on the first time step, $w_{\mathrm{H2–B2}}$ remains equal to zero. In other words, after one time step, the first inequality in (A4) will become true. By symmetry, the remaining (A4) inequalities will also become true on this time step. Thus, the base case is established.

A1.2 Induction step

We examine the derivative of the error function with respect to the weight $w_{\mathrm{H2\text{–}B}k}$ $(1 < k < n)$ when (A1)–(A3) hold. We will show that when $w_{\mathrm{H2\text{–}B}k} = w_{\mathrm{H2\text{–}B}(k-1)}$ this partial derivative is positive while when $w_{\mathrm{H2\text{–}B}k} = w_{\mathrm{H2\text{–}B}(k+1)}$ this partial derivative is negative. Analogous conditions hold for $w_{\mathrm{H2\text{–}B}(n-k+1)}$, $w_{\mathrm{H2\text{–}D}k}$, and $w_{\mathrm{H2\text{–}D}(n-k+1)}$.

The first derivative of the (cross entropy) error function with respect to input weight w_{ij} on presentation of a particular pattern p is given by

$$\frac{\partial E_p}{\partial w_{ij}} = -f'(\mathrm{net}_{pi})\left(\sum_k w_{ki}\delta_{pk}\right)a_{pj},$$

where net_{pi} is the net input to (hidden) unit i, $f(x)$ is the hyperbolic tangent function, w_{ki} is the weight from hidden unit i to output unit k, $\delta_{pk} = t_{pk} - a_{pk}$ where t_{pk} is the target value for output unit k on pattern p, a_{pk} is the activation of output unit k, and a_{pj} is the activation of input unit j (Rumelhart, Durbin, Golden, & Chauvin, 1995).

Suppose $w_{\mathrm{H2\text{–}BK}} = w_{\mathrm{H2\text{–}B}(k-1)}$. The two relevant patterns are $p = \mathrm{DB}_{k-1}$ and $p = \mathrm{BD}_k$. We consider each of these in turn, dropping the p subscript since we are focusing on one pattern at a time:

- $p = \mathrm{DB}_{k-1}$:

  Target $= [0, 1, 1, 0]$

  $$\mathrm{net}_{\mathrm{H2}} = w_{\mathrm{H2-D}(k-1)} + w_{\mathrm{H2-B}k} = 0$$

  since $w_{\mathrm{H2-B}k} = w_{\mathrm{H2-B}(k-1)} = -w_{\mathrm{H2-D}(k-1)}$ by (A3)
  Thus,

  $$f'(\mathrm{net}_{\mathrm{H2}}) = f'(0) = \frac{1}{2}(1+0)(1-0) = \frac{1}{2}$$
  $$\sum_k w_{ki}\delta_{pk} = w_{\mathrm{oh}}(0 - \mathrm{sig}(\mathrm{net}_{\mathrm{o1}})) - w_{\mathrm{oh}}(1 - \mathrm{sig}(\mathrm{net}_{\mathrm{o2}})) - w_{\mathrm{oh}}(1 - \mathrm{sig}(\mathrm{net}_{\mathrm{o3}})) + w_{\mathrm{oh}}(0 - \mathrm{sig}(\mathrm{net}_{\mathrm{o4}})).$$

  Since $\mathrm{net}_{\mathrm{H1}} = 0$ as well (because the D and B inputs to H1 are balanced), $\mathrm{sig}(\mathrm{net}_{Ok}) = ½$ for $k = 1, 2, 3, 4$. Therefore,

  $$\sum_k w_{ki}\delta_{pk} = w_{\mathrm{oh}}\left(0 - \frac{1}{2}\right) - w_{\mathrm{oh}}\left(1 - \frac{1}{2}\right) - w_{\mathrm{oh}}\left(1 - \frac{1}{2}\right) + w_{\mathrm{oh}}\left(0 - \frac{1}{2}\right) = -2w_{\mathrm{oh}}.$$

  Since $a_{Bk} = 1$,

  $$\frac{\partial E_{\mathrm{DB}_{k-1}}}{\partial w_{\mathrm{H2-B}k}} = -\frac{1}{2}(-2w_{\mathrm{oh}}) = w_{\mathrm{oh}}$$

- $p = \mathrm{BD_k}$:

  Target $= [1, 0, 0, 1]$

  $$\mathrm{net}_{\mathrm{H2}} = w_{\mathrm{H2-B}k} + w_{\mathrm{H2-D}(k+1)} \leq w_{\mathrm{H2-B}k}$$

  since $w_{\mathrm{H2-B}k} = w_{\mathrm{H2-B}(k-1)} > -w_{\mathrm{H2-D}(k+1)}$ by (A3)

Moreover, since $w_{\mathrm{H2-D}(k+1)} \leq 0$ and $w_{\mathrm{H2-B}k} > 0$ (by (A2)–(A3)), $\mathrm{net}_{\mathrm{H2}} > 0$. Thus

$$f'(\mathrm{net}_{\mathrm{H2}}) < \frac{1}{2}.$$

Let $\varepsilon = f(\mathrm{net}_{\mathrm{H2}})$. Then $\varepsilon > 0$. As above, $\mathrm{net}_{\mathrm{H1}} = 0$, so

$$\sum_k w_{ki}\delta_{pk} = w_{\mathrm{oh}}(1 - \mathrm{sig}(w_{\mathrm{oh}}\varepsilon)) - w_{\mathrm{oh}}(0 - \mathrm{sig}(-w_{\mathrm{oh}}\varepsilon)) - w_{\mathrm{oh}}(0 - \mathrm{sig}(-w_{\mathrm{oh}}\varepsilon)) + w_{\mathrm{oh}}(1 - \mathrm{sig}(w_{\mathrm{oh}}\varepsilon)),$$

where $\mathrm{sig}(x) = \frac{1}{1+\mathrm{e}^{-x}}$ is the logistic function.

Thus,

$$\sum_k w_{ki}\delta_{pk} = 4\left(\frac{1}{2}^{-}\right) w_{\mathrm{oh}} < 2w_{\mathrm{oh}},$$

where $\frac{1}{2}^{-} = \mathrm{sig}(-w_{\mathrm{oh}}\varepsilon) = 1 - \mathrm{sig}(w_{\mathrm{oh}}\varepsilon)$.

Again, $a_{\mathrm{BK}} = 1$, so

$$\frac{\partial E_{\mathrm{BD}_k}}{\partial w_{\mathrm{H2-B}k}} = -xy.$$

Where ½ $> x > 0$ and $2w_{\mathrm{oh}} > y > 0$.

Therefore,

$$\frac{\partial E_{\mathrm{BD}_k}}{\partial w_{\mathrm{H2-B}k}} = w_{\mathrm{oh}}^{-}.$$

Where $w_{\mathrm{oh}} > w_{\mathrm{oh}}^{-} > 0$. Putting these two values together yields:

$$\frac{\partial E}{\partial w_{\mathrm{H2-B}k}} = w_{\mathrm{oh}} - w_{\mathrm{oh}}^{-} > 0.$$

An analogous argument shows that when $w_{\mathrm{H2-BK}} = w_{\mathrm{H2-B}(k+1)}$,

$$\frac{\partial E}{\partial w_{\mathrm{H2-B}k}} < 0$$

as desired.

Appendix 2

Here we present one of the solutions for the Ring input. We used the same three layer feedforward model as Basic Model. Instead of 2 hidden nodes, there are four hidden nodes in this solution. The first two hidden units are similar to those of the Idealised Network. Here we assume that the weights from the input units to the first two hidden units are all positive for B units and all negative for D units. However, the absolute values increase monotonically with position for the first hidden unit and decrease monotonically with position for the second hidden unit. The third hidden unit detects the input, D1B10. The fourth hidden unit detects the input, B1D10.

Table A1. Weights from input to hidden units (columns index inputs, rows index hiddens).

		1	2	3	4	5	6	. . .	19	20
	Bias	B1	D1	B2	D2	B3	D3		B10	D10
H1	0	1	−1	2	−2	3	−3		10	−10
H2	0	10	−10	9	−9	8	−8		1	−1
H3	−12	7	0	0	0	0	0		0	7
H4	−12	0	7	0	0	0	0		7	0

Table A2. Weights from hidden to output units (columns index hidden units, rows index outputs).

	Bias	H1	H2	H3	H4
1 B1	−4	0	10	−20	20
2 D1	4	0	−10	20	−20
3 B2	−4	10	0	20	−20
4 D2	4	−10	0	−20	20